Graphical Tools for Item Response Theory Model Assessment

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon

By
Jian'ou Zhang

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Mathematics and Statistics
142 Mcclean Hall, 106 Wiggins Road
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5E6
Canada
OR
Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

Item response theory(IRT) is widely used in many fields such as psychology, education and health. IRT model assessment is essential because model-data misfit can result in the risk of drawing incorrect inferences and conclusions. There have been extensive work on model assessment for item responses theory, but most literature mainly concentrates on theoretical methods such as test statistic procedures for goodness-of-fit. Though graphical diagnosis tools have been explored in the current literature, it is still not enough and needs more work. Hence, our work focus on exploring graphical diagnosis tools for assessing model fit in IRT contexts. First, we compare the observed and expected sum scores through plot. Second, we propose residual diagnostic plots based on randomized quantile residual(RQR). Finally, we consider comparing a non-parametric model fit with the posited parametric model fit via item characteristic curves(ICC). The first method has been long recognized in the existing literature, while the remaining two methods are proposed and new in this thesis, which is actually a contribution of my research. Also, in each of methods, We consider both in-sample and out-of-sample prediction. A simulation study has been conducted to evaluate and compare the performance of these methods. Our preliminary results indicate that observed v.s expected sum scores fails to detect lack of model fit. For RQR checking, out-of-sample prediction outperforms in-sample prediction in terms of detecting the misfit, while non-parametric methods seem to be promising for model assessment of a parametric model.

Acknowledgements

I would very much like to appreciate the instructions and helps of my supervisor, Dr Juxin Liu, sincerely. Thanks for her carefully advice, suggestions and encouragement in instructing me on my thesis work, as well as the financial support she provides to me. All these helps enable me to focus on and complete my research with high efficiency. Also, my supervisor's noble characteristic such as confidence and hardworking have had a profound impact on me, making me be still optimistic when I faced difficulties.

I really give my acknowledgement to my committee members, Dr Steven Rayan, Dr. Shahedul Khan and Dr. Zhi Li. Thanks for their help as well as comments on my thesis improvement. Thanks to Huokai Wu and Professor Derek Postnikoff in my research team to give me opinions on my research.

As well, I am very grateful to the Department of Mathematics and Statistics of University of Saskatchewan in providing me assistantship. Finally, all my love and thanks to those professors and students in the Department of Mathematics and Statistics of University of Saskatchewan for their help on both my academic work and life.

Especially thanks to my roommates Yixiao Gao and Huiyao Kuang, and my family members for their help, understanding and support on my work.

Table of Contents

List of Figures

List of Abbreviations

IRT	Item Response Theory
1PL	One-parameter Logistic Model
2PL	Two-parameter Logistic Model
3PL	Three-parameter Logistic Model
MCMC	Markov Chain Monte Carlo
IRF	Item Response Function
ICC	Item Characteristic Curve
CDF	Cumulative Distribution Function
PDF	Probability Density Function
PMF	Probability Mass Function
RQR	Randomized Quantile Residual
LM	Lagrange Multiplier
TOEFL	Test of English as a Foreign Language
PISA	Programme for International Student Assessment
PPMC	Posterior Predictive Model Checking
GR	Graded Response Model
MPT	Math Placement Test
MLE	Maximum Likelihood Estimate
GRE	Graduate Record Examinations

1. Introduction

Item Response Theory (IRT) is a family of statistical models which are used to characterize the relationship between a test taker's performance on a given item and the test taker's latent traits (unobservable characteristics, such as mathematical ability) being assessed by this given item. An Item Response Function (IRF) is used to mathematically model such relationship. An Item Characteristic Curve (ICC) is the visualization of an IRF. For example, the following is an IRF with respect to a given item:

$$P(\text{a person of trait } \theta \text{ correctly answers an item of difficulty } b) = \frac{\exp(\theta - b)}{1 + \exp(\theta - b)}. \quad (1.1)$$

Also, the corresponding ICC of equation(1.1) is displayed in Figure 1.1 (https://hansjoerg.me/2018/04/23/rasch-in-r-tutorial/).

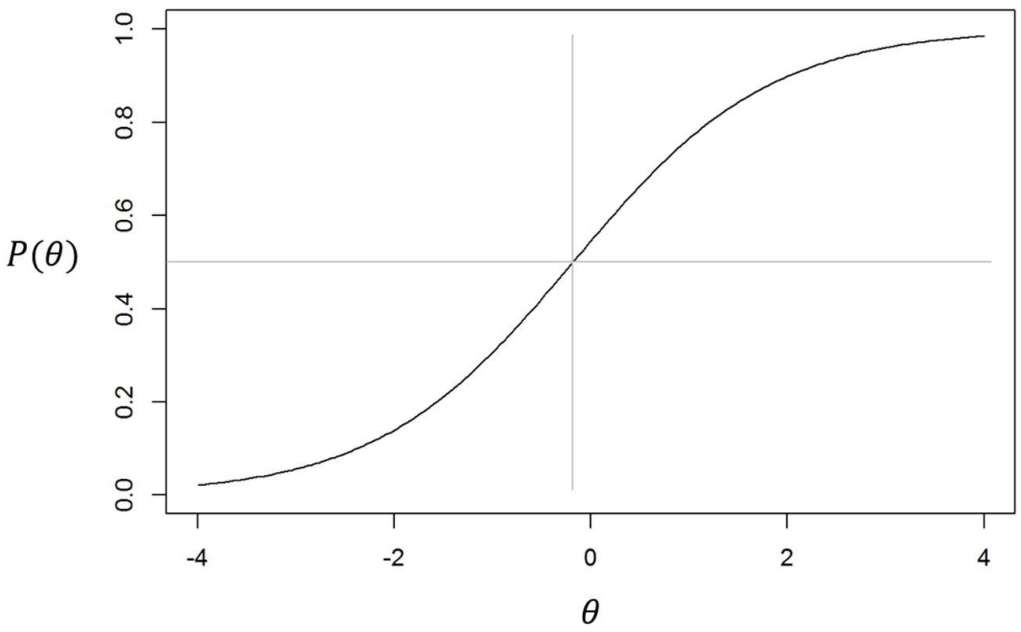


Figure 1.1: ICC of Rasch Model.

Figure 1.1 actually reflects the relationship described by equation (1.1). IRT models are widely used in a large variety of fields such as psychology, education and health. For example, the Test of English as a Foreign Language (TOEFL) and Programme for International Student Assessment (PISA) test use IRT models to assess test takers' latent traits. The

most commonly used parametric IRT models are Rasch model, two-parameter logistic model (2PL) and three-parameter logistic model (3PL). One thing that we need to mention is that Rasch model is not identical to one-parameter logistic model (1PL) and the details could be found on Columbia University Mailman School of Public Health (https://www.publichealth.columbia.edu/research/population-health-methods/item-response-theory). There are also other parametric IRT models, the reason why we only consider the above three IRT models is that they are the most commonly used and simplest IRT models. Also, there are several IRT model assumptions. The detailed explanation of IRT models and model assumptions will be given in the next chapter.

IRT models has some advantages, which enables it to be very crucial in real applications. For example, Hambleton et al.[24] pointed out that one advantage of IRT models is that it permits test designers to design some tests with particular characteristics for some specific test takers, e.g., a test designed to select high ability test takers for scholarship. Also, for test designers, if the analysis for the responses to an item shows absurd information, e.g., the decreasing ICC, it means that this item is problematic, and this problematic item will be removed from test.

IRT model checking is a very vital step before we draw conclusions. Sinharay[47] indicated that ignorance of model assessment could be under the risk of drawing incorrect conclusion. Swaminathan[53] demonstrated that since item response theory relies on strong mathematical and statistical assumptions, only these assumptions are met could IRT models be used effectively to analyze the data and draw inferences. Hence, implementing model checking for IRT models is essential and the purpose is to check whether the fitted model is able to explain the data adequately. IRT models can be classified as unidimensional and multi-dimensional cases, where the former only involves one latent trait and the latter has two or more latent traits. On the other hand, the responses to items could be binary (response only has two possibilities, e.g. correct/incorrect, yes/no, true/false) or polytomous (response will have more possibilities, for instance, rating a service on a 5 point Likert scale). In this thesis, we will simulate unidimensional latent trait with binary responses items from IRT model and we have 526 test takers and 30 items in the simulation.

Assessing IRT model can be tackled in two directions:[53] (1) checking model assump-

tions, that is, whether model assumptions are violated or not (Yen[57], Stout[52]). (2) checking model predictions, meaning that whether the model-predicted values are similar to the observed values or not (Sinharay[47], Bryonna[5]). Assessing each of these two directions can be done by assessing different model aspects: item fit (Bock[4], Yen[56], Stone and Zhang[49], Kyong et al.[28]), person fit (Glas[14], Glas et al.[17], Ferrando[10]) and overall fit (Sinharay[47], Alberto[34]). The assessing tools can be goodness-of-fit test statistics and diagnosis graphs and details will be given in the following. It must be pointed out here that though there's no agreement that which method is the best for IRT model assessment, it is suggested that both test statistic and graphical diagnosis should be used together for model checking (Liang[31]).

Goodness-of-fit test statistics describe the differences between observation and model-predictions and are frequently used to assess IRT models. There are some goodness-of-fit test statistics. For instance, Bock's chi-square test[4], Yen's Q_1 test[56] and likelihood ratio test G^2 (McKinley and Mills[37]). But several problems exist for applying them to proceed model checking. Hambleton et al.[22] indicated that these traditional test statistics are sensitive to the number of ability groups. This means that when the number of ability group is large, statistical test is able to identify very small discrepancy between observed and model predicted values. Thus, hypothesis testing would become nonsensical in this case since the null hypothesis will always be rejected (Liang[31]). The second drawback of above test statistics is that they do not approximately follow Pearson chi-square distribution. Stone and Zhang[49] demonstrated that one reason is the fact that test statistic followed chi-square distribution is based on the true values of theta (latent trait), while the estimates of theta from IRT models are treated as true values in calculating the test statistic, and since the estimates contains errors, this could make the test statistic fail to follow chi-square distribution. Another reason is that the degree of freedom of chi-square tests is in question. Orlando and Thissen[39] pointed out that since the calculation of test statistic involves model-dependent latent abilities, the degree of freedom might not be as what Yen[56] and McKinley and Mills[37] claimed and the simulation study of Orlando and Thissen[39] showed this. Hence, Orlando and Thissen[39] defined a new Pearson chi-square test statistic $S - X^2$ and a new likelihood ratio statistic $S - G^2$. The advantage of these two test statistics is that

the latent traits estimations only depend on observed data rather than models. Stone[48] proposed X^{2*} and G^{2*} statistics, which both involve posterior expectation of latent traits instead of point estimates in order to consider the uncertainty of latent traits. These improved test statistics are found to outperform previous test statistics for assessing some IRT models (Stone and Zhang[49], Von et al.[46], Kang and Chen[26]). Glas[15][16] developed Lagrange Multiplier (LM) taking into account uncertainty of item parameters and Glas and Falcon[18] applied LM to IRT model assessment. Their simulation study shows that LM procedure is much better than Yen's Q_1 test, but the overall characteristic (i.e., Type I error, power and false positive rates) of $S - X^2$ is better than that of LM. Maydeu-Olivares and Joe[33] indicated that goodness-of-fit assessment in binary response IRT models is actually assessing a 2^I contingency table, where I is the number of items. However, when I becomes large, the number of cells in the contingency table increases exponentially and this would lead to sparseness of contingency table (many cells in the contingency table have very small value or zero). This sparseness could cause the difficulty in parameter estimation. Also, residual check does not work in this case because it's hard to find the trend of residuals which is used for checking model misfit. So they defined two classes of quadratic forms of limited information test statistics L_r and M_r. L_r and M_r are both on the basis of marginal residuals up to the order r (low-dimensional residuals) of high-dimensional contingency table, but L_r uses known model parameters while M_r uses estimated model parameters in light of Maximum Likelihood Estimate (MLE). Maydeu-Olivares and Joe[33] showed that both L_r and M_r converges to the chi-square distribution under null hypothesis, and marginal residuals could be effective to detect the lack of model fit. Swaminathan[53] pointed out that limited information test statistics is a contribution for testing goodness-of-fit 2^I contingency table and the limitation is it could only apply to binary response.

Graphical diagnosis of checking IRT model misfit has been explored in the literature as well. It is very intuitive and strongly recommended by Hambleton and Rogers[23]. Graphical checking could be executed by plotting observed and expected values directly and doing comparison. For example, Hambleton, Swaminathan and Rogers[23] assumed test takers' latent traits follows standard normal distribution and then they binned test takers based on test takers' sum scores (test takers' sum scores were ranked first, then all the test takers were

put into different bins in light of the rank of their sum scores). They calculated the observed proportion of correctly response of every bin with respect to each item and compared these proportions to the corresponding estimated ICCs of one-parameter, two-parameter and three-parameter logistic models through graphs. Finally obvious discrepancy was found when the model did not fit the observed data well. Kalinowski[25] extended Hambleton's work and developed some improvements. For instance, Hambleton used midpoint of each bin as the estimated latent trait to calculate the expected probability of responding correctly to each given item for that bin. Kalinowski indicated that since latent trait is assumed to follow standard normal distribution, usually there are more test takers with latent traits at one end than the other end of the bin, meaning that using midpoint could not well reflect the feature of the test takers' latent trait in the bin and hence fails to give a good estimation of the expected success probability for each bin. Also, Kalinowski illustrated that latent trait changes across the range of the bin, resulting in the corresponding changes of expected success probability on that bin, it turns out that the expected success probability for each bin based on midpoint did not take into account such changes. Hence, Kalinowski instead used an integral on each bin to represent expected success probability for that bin and the corresponding bin plots he showed do an excellent job in detecting model misfit. Swaminathan[53] compared observed and expected test score distribution from one-parameter, two-parameter and three-parameter logistic models by graphs and succeeded in identifying the model which misfits the data. However, there is a problem. Sum scores collapse the information of test takers' responses to items (e.g., the sum scores fail to display how each test taker responds to a particular item), but these information might reveal the model misfit (Alberto Maydeu-Olivares[34]). So, when the discrepancy between observed sum scores and expected sum scores is very small, it's not always safe for us to conclude the model fits the data well. And this is in fact what happens in our simulation study (section 4.2.1). We compare observed and expected item sum score from all candidate models via plots. Unfortunately, the difference between observed values and expected values from wrong models under some case is really small. Put it in another way, comparing observed and expected values directly for detecting model-data misfit might not always work. Graphical diagnosis has been conducted based on residuals analysis as well. Residuals are calculated as the differences between observed and model-

predicted values. Generally, residuals are normalized by dividing their estimated standard error at each predicted value. This normalization will make the residuals have the same scale so they could be compared. The normalized residual is called *standardized residual*. Residuals could be plotted to assess model goodness-of-fit. Usually, residuals are put in the vertical axis and the horizontal axis is an independent variable. If the model fits the data well, the residuals from this fitted model should be randomly distributed, namely, there is no obvious pattern in the residual plot. Hambleton, Swaminathan and Rogers[23] applied residual plots in checking 1PL, 2PL and 3PL IRT model for fitting some given binary responses. They calculated standardized residuals by

$$z_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{\frac{E_{ij}(1-E_{ij})}{N_j}}}. \tag{1.2}$$

where O_{ij} is the observed proportion of responding correctly for j-th latent trait bin with respect to the i-th item, E_{ij} is the expected proportion of obtaining correct response to i-th item for j-th latent trait bin from fitted model and N_j is the number of test takers in j-th latent trait bin. Then they put standardized residuals from each fitted model in a plot with x-axis standing for latent trait to see whether any clear pattern could be found. The result indicated that apparent pattern exists in the residual plot of 1PL model, and 3PL model residuals is most random, meaning that 3PL fits the given data best. In fact, Hambleton and Swaminathan[22] and McDonald[36] demonstrated that incorporating residual analysis would be most powerful in assessing model-data fit. However, Bryonna Bowen[5] pointed out that sometimes, there still can be some pattern in the residual plots even if the model fits data well and this casts the doubt of usefulness of identifying model misfit by checking residual plots. To determine whether a residual plot really worked in assessing goodness of model-data fit, she utilized three simulations generated from 3PL model and fitted the 3PL model (true model) and Rasch model (wrong model), then calculated residuals through some improved techniques and finally generated corresponding residual plots for comparison. The number of simulations is three rather than one because this can verify that if some pattern appears in the residual plots for all three simulations, then this pattern is not random by chance. Her result showed that residuals from both models displays patterns and there is no noticeable difference between these patterns, but the magnitude of residuals from Rasch

model is much larger than that of 3PL model. It seems that in light of her simulation, we cannot only depend on whether there is a pattern or not in the residual plot to identify model misfit. This is actually a flaw of graphical diagnosis: it is less objective than test statistics in judging whether the model fits data well or not, but graphs (e.g., residual plots) can still provide meaningful insight for identifying model-data misfit (Liang[31]). Anyway, Bryonna's simulation study provides a contribution in residual graphical diagnosis since there is little concern on using residual analysis to assess model-data misfit in IRT framework (Sinharay[47]).

Posterior Predictive Model Checking (PPMC) method is a Bayesian method for assessing model-data misfit. The Bayesian method treats model parameters as random variables rather than fixed values. Adding randomness of parameters in the model allows the model to be more flexible and give better predictions from the model (Rubin[43]). Also, Bayesian method brings large convenience in computing many complicated integrals in statistics (Rubin[43]). PPMC now successfully attracts great attentions in IRT model assessment because it is simple, intuitive and has strong theoretical basis (Sinharay[47]). The main idea of PPMC is to contrast observed data to model-predicted data under appropriate test statistics. The PPMC is calculated as follows (Sinharay[47]):

$$p(y^{rep}|y) = \int p(y^{rep}|\omega)p(\omega|y)d\omega. \tag{1.3}$$

where y^{rep} is the replicated data, y is the observed data, $p(y^{rep}|y)$ is the posterior predictive distribution of y^{rep}, ω is a vector representing model parameters. $p(\omega|y)$ is the posterior distribution of ω based on y. $p(y^{rep}|\omega)$ is the likelihood function of a distribution. Then a test statistic $T(y)$ is selected in order to measure the discrepancy of model-data fit. If significant discrepancy exists between $T(y)$ and $T(y^{rep})$, it demonstrates the model failure. Basically, such discrepancy could be indicated by *posterior predictive p-value* (PPP-value). and it is actually the Bayesian counterpart of the classical p-value (Sinharay[47]). The formula for PPP-value is(Sinharay[47]):

$$p = P(T(Y^{rep}) \geq T(y)|y) = \int_{T(Y^{rep})\geq T(y)} P(Y^{rep}|y)dy^{rep}. \tag{1.4}$$

From above equation, we can see that PPP-value is the probability that the replicated data is more extreme than the observed data under the test statistic $T(y)$. In practice, it is very hard

to compute the exact value of this integral, especially when the number of parameters is huge. Notice that this integral is in fact the expectation of the posterior predictive distribution of Y^{rep} on the set such that $T(Y^{rep}) \geq T(y)$, hence, by Central Limit Theorem, we are able to approximate this expectation by sample mean. The steps are in the following (Gelman et al.[13])

1. Draw N samples from posterior distribution of parameter ω, we denote these N samples by $\omega_1, ..., \omega_N$.

2. For each ω_i, we draw a y_i^{rep} from the joint posterior distribution $P(Y^{rep}, \omega|y) = P(Y^{rep}|\omega)P(\omega|y)$.

3. Calculate the test statistic $T(y, \omega_i)$ and $T(y_i^{rep}, \omega_i)$ for each i.

4. Calculate the proportion such that $T(y, \omega_i) \geq T(y_i^{rep}, \omega_i)$ and this proportion is the estimated PPP-value.

if PPP-value is close to 0 or 1, usually it indicates model-data misfit, namely, the observed data and posterior predictive data are significantly different. While close to 0.5 demonstrating model-data fit. However, it is found that PPP-value was closer to 0.5 more often than what was expected based on uniform distribution (Levy[29]). So applying PPP-values would sometimes result in conservative inferences (Fu,J., Bolt, D. M., and Li, Y.[11], Sinharay[47]).

Guttman[19] introduced PPMC, Rubin[43] developed formal definition of PPMC, and Sinharay[47] illustrated that PPMC could be able to be used for assessing different aspects of IRT models (e.g., item fit, person fit, overall fit) and also, PPMC can be run by graphical plots or PPP-values. For example, Sinharay[47] applied PPMC to assess whether a simple 3PL model or a more complicated hierarchical model could adequately explain the data consisted of sixteen item models covering four main content areas with four difficulty levels: very easy, easy, hard and very hard. The difference between two models is that 3PL model takes no account of variation between items within a model. The test statistic which Sinharay used is standard deviations of observed data and model predicted data. He displayed these standard deviations through boxplots to do comparison. The result showed that hierarchical model works better than 3PL model for fitting the given data. Zhu[58] utilized PPMC with

simulations to check the IRT model assumptions. The simulation is responses generated from a simple-structure two-dimensional polytomous graded response model (GR model) with local dependence. The unidimensional GR model is used to fit the simulation and PPMC is considered to identify violation of unidimensionality and local independence. Different test statistics are used by the author: *total test score distribution of observed vs. posterior predictive data*(test-level test statistic), *item score distribution, item-total score correlation, Yen's Q_3[55], Stone's item-fit test statistic[48]*(these four are considered as item-level test statistics) and *global odds ratio, Yen's Q_3 and absolute item covariance residual*(these three are used as pair-wise statistics). PPP-values for each test statistic show that pair-wise test statics are more powerful in detecting violation of unidimensionality and local independence than test-level and item-level ones, which is consistent with those finding of Levy[29]. Zhu[58] indicated the reason is that no parameter in the unidimensional GR model describes the associations between responses to pairs of items, but the pair-wise test statistics could detect these associations. Zhu's simulation results as well shows that among all test statistics, Yen's Q_3 performs best, while total test score distribution and item score distribution the worst. This also emphasizes the importance of selecting test statistic for PPMC. There are other works in application of PPMC to assess IRT model-data fit(Li et al.[30], Kuhfeld[27]), and the results verify the usefulness of identifying IRT model misfit by PPMC.

Most IRT models are parametric models, namely, the model involves model parameters (e.g., item difficulty). However, these parametric IRT models rely too much on model assumptions. When these assumptions are met, parametric IRT models could be applied in practical situations (Swaminathan[53]). Therefore, non-parametric IRT models are introduced because of the advantage that they are based on minimum assumptions (Van der Linden[54]). The non-parametric IRT models was introduced first by Guttman[20][50][51] and gained development in psychology (Mokken[54], Molenaar[54], Ramsay[54][40][42]) in the past few decades. Since less assumptions are needed than parametric IRT models, non-parametric IRFs are closer to the 'true IRF' than parametric IRFs, so they serve as an effective tool for assessing parametric model-data misfit (Van der Linden[54]). By Douglas and Cohen[7], noticeable difference between estimated non-parametric model and parametric model indicating parametric model failure. There has some research in assessing parametric

model by using non-parametric models (Douglas and Cohen[7], Liang[31]), however, it's very limited. More work for taking into account non-parametric model to check lack of parametric model fit is needed.

In my thesis, we will develop graphical tools for checking the fit of given IRT models by item fit. Our data is a simulated binary response pattern matrix and candidate models are Rasch model, 2PL model and 3PL model. As well, there are two cases, the in-sample and the out-of-sample cases. In-sample means that the entire simulated response pattern matrix is used to fit all models, while out-of-sample comes from the idea of simple cross validation, i.e, we randomly separate the matrix into training data set and validation data set. Three graphing methods are considered here. The first one is observed v.s expected item sum score plot. This method is very straightforward and has been utilized before, but it is shown that it does not work well in assessing goodness-of-fit. The second method is called *Randomized quantile residual (RQR)*. RQR was proposed by Dunn and Smyth[8] for handling discrete observations and has been successfully applied to assess bio-statistical model-data fit (e.g., non-normal regression model[45], generalized linear mixed model[2]), however, there's no such work in IRT model assessment framework. Hence, what we are going to do in our thesis is to generate RQRs for each item on the basis of test takers' estimated success probabilities with respect to this item. We do this process for each candidate model. Then RQRs for each item of each model will be displayed and contrasted with standard normal distribution through Q-Q plot. We will prove in chapter 3 that RQRs from true model approximately follow standard normal distribution. Therefore, Obvious difference between RQRs quantiles and theoretical quantiles indicates model misfit. The simulation study shows that RQRs of all items for in-sample case fail to detect model misfit, while for the out-of-sample, RQRs do identify the inadequacy of some fitted model via several items. We conclude that the RQRs for out-of-sample performs better than in-sample case in detecting model misfit. This is actually a contribution of this thesis. The third checking method of my thesis is the consideration of kernel smooth checking method, a non-parametric way proposed by Ramsay[41], for assessing the model-data fit of a given parametric model. An advantage of the kernel smooth is that fewer model assumptions are required. The essential idea is to compare a non-parametric model fit with the posited parametric model fit via item characteristic curves. According

to our result, the non-parametric does work well in assessing parametric model-data misfit. Hence, this is consistent with the claim of Van der Linden and Hambleton[54]: the non-parametric model is promising for assessing the fit of parametric IRT models.

The remaining part of the thesis is organized as follows. Chapter two gives a detailed introduction to three parametric IRT models and model assumptions as well as nonparametric model. In chapter three, we present three model assessment methodologies. Next we apply these methods in our simulation study and display the result in chapter four. Finally, conclusions as well as future work are discussed in the chapter five.

2. IRT Models

In this chapter, we will introduce three parametric models: Rasch model, Two-parameter model (2PL) and Three-parameter model (3PL). Since these parametric IRT models requires several assumptions, we list them in the following:

- Dimensionality Assumption: Dimensionality means the number of latent traits. Since IRT models involve person's latent traits, some models contain only a single latent trait, which we call unidimensional IRT. For example, the Rasch, 2PL and 3PL models in this thesis are all unidimensional. There are others IRT models enclosing two or more latent traits and these are multi-dimensional IRT. For instance, R.J.de Ayala[1] shows a two-dimensional IRT model involving both math and reading abilities.

- Local Independence: This means that a correct or incorrect response to one item does not result in correct or incorrect response to any other items.

- Monotonicity: This implies that a person with a higher latent trait level is more likely to correctly respond to the given item.

Before applying parametric IRT models to the data, one needs to check carefully that whether all the assumptions are satisfied. Any violation of these assumptions might result in inaccurate outcomes of utilizing parametric IRT models to fit the given data.

2.1 Rasch Model

Rasch Model is a parametric IRT model which relates a person's latent ability to an item parameter, i.e., item difficulty, through a simple mathematical form:

$$P(X_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}. \tag{2.1}$$

where

- X_{ij} for the response of i-th person with respect to the j-th item. It is binary, that is, the value of X_{ij} is 0 or 1. $X_{ij} = 1$ means that the i-th person responds correctly to the j-th item, and $X_{ij} = 0$ indicates incorrect response for the i-th person to j-th item.

- θ_i for the i-th person's latent ability.

- β_j for the j-th item difficulty.

From above, we can see that the left hand side is the probability of i-th person responding correctly to the j-th item if his/her latent ability is θ_i and the item difficulty is β_j, and this probability equals to the right hand side. Hence, for a fixed latent trait, the more difficulty the item is, the lower the probability for obtaining the correct response, and similarly, for fixed item difficulty, the higher the latent ability, the larger the probability to respond correctly to the given item. One thing we need to mention here is that Rasch model is a special form of One-Parametric Logistic model (1PL). For 1PL model, all the items have different item difficulties but same item discrimination α. In Rasch model, all item difficulties equal to 1, while in general 1PL, this item discrimination can be other numbers. We will further explain item discrimination in 2PL model.

Usually, in most IRT models, it is assumed that $\theta \sim N(0, 1)$, meaning that the distribution of latent ability follows standard normal distribution. Thus, $\theta > 0$ implies the test taker's latent ability level is above average, and, accordingly, negative value of θ means it is below average. However, as Ramsay[40] points out, if $h(\theta)$ is a differentiable and strictly monotonic function of θ, then there exists a set of functions P_j^* such that $P_j(\theta) = P_j^*(h(\theta))$[40]. This indicates that the scale of θ is transformable. Based on this fact, we can transform the range of θ to a bounded interval. For example, Ramsay[40] defined *Tiltted Scaled β Distribution (TSβ)* for latent ability. The measurement of θ in this distribution is between 0 and some fixed upper limit $T(\theta \in [0, T])$. The benefit of TSβ distribution is that it could help identifying ICCs uniquely[40]. In our thesis, we still assume that the distribution of θ is standard normal because it is the most commonly used distribution for latent traits. Notice that the scale of θ is $(-\infty, \infty)$, but in application, $\theta \in (-3, 3)$ usually.

Equation (2.1) is the IRF of Rasch model. As we explained in chapter one, an IRF mathematically characterizes the relationship of latent traits and probabilities of endorsing

an item. The equation (2.2) and (2.3) in the following are IRFs of 2PL and 3PL respectively, which are more complicated because these two models involve more item parameters. If we display an IRF through graph, we obtain an ICC, that is, ICC reflects the relationship characterized by IRF graphically. An ICC is plotted in xOy plane, where x-axis denotes the latent trait and y-axis the probability of responding correctly to the given item. Generally, an ICC is increasing, indicating that the higher the latent trait, the larger the probability of endorsing the given item. However, sometimes, an ICC is decreasing. This means the model assumption of monotonicity is violated and the given item is problematic.

2.2 2PL Model

The Two-parameter Logistic model (2PL) involves person's latent ability and two parameters– item discrimination and item difficulty. The mathematical expression is:

$$P(X_{ij} = 1|\alpha_j, \beta_j, \theta_i) = \frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j)))}. \tag{2.2}$$

α_j is called j-th item discrimination, which is non-negative. It denotes the ability for item j to distinguish person with different latent abilities. Generally, the larger the item discrimination, the easier the item could discriminate high-ability from low-ability person. However, by Baker[3], if $\alpha_j > 1.7$, then the ability for j-th item to distinguish test takers will be very high. In our thesis, $\alpha_j \in (0, 3)$ for each j.

Reflected on the ICCs, item discriminations affect curve steepness. The larger the α_j, the steeper the ICC of j-th item. This means even if two test takers' latent abilities are very close, there's still obvious difference between their probabilities of endorsing j-th item.

Note that if we take all the item discriminations in the 2PL equal to some non-negative constant, i.e. $\alpha_j = \alpha$ for all j where $\alpha \geq 0$ is a constant, it turns out that this model becomes 1PL model, and Rasch model if $\alpha = 1.0$. Hence, 1PL is a special case of 2PL model.

In practice, 2PL is more common than 1PL, since item discrimination is allowed to be different for the former. In many actual standard exams, such as GRE (Graduate Record Examinations) or TOEFL test, it is impossible for all items to have the same item discrimination, there are always some items which are able to distinguish person latent abilities better than other items.

2.3 3PL Model

Comparing to 2PL model, Three-parameter Logistic model (3PL) adds one more parameter in the model. This new parameter is called *pseudo-guessing parameter*. So 3PL model adds the possibility of obtaining the correct response only due to chance. We denote this parameter by γ_j for the j-th item. Hence, the model is:

$$P(X_{ij} = 1|\alpha_j, \beta_j, \gamma_j, \theta_i) = \gamma_j + (1 - \gamma_j)\frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))}. \tag{2.3}$$

the pseudo-guessing parameter γ_j is the probability of responding correctly to the j-th item when the test taker's latent ability is super low, indicating the test taker's success only due to chance. In fact, R.J.de Ayala[1] pointed out that γ_j is a pseudo-guessing instead of guessing parameter because it is lower than what is predicted from random guessing model. By Hambleton[23], the reason is that for multiple-choice items, item designers develop many attractive but incorrect choices, which makes test takers be more likely to select them. But for random guessing model, each choice is assumed to be selected with equal chance by test takers.

Also, γ_j is a positive number, and if we look at the item characteristic curve, it is where the left asymptote approximates as latent ability goes to negative infinity (i.e. the latent ability level is very very low), while for Rasch and 2PL model, the left asymptote are both zero, this is because in these two models, we think test takers select the correct response since they really know it instead of guessing, and if a test taker's latent ability is extremely low, he or she will choose the incorrect response, meaning that there's no chance to obtain the correct response.

In 3PL model, if $\gamma_j = 0$ for all items, we obtain 2PL model. Up to now, we can see these three models get forward one by one: Rasch the simplest, then 2PL, and 3PL the most complex. However, it does not mean that 3PL model will always fits the data best in real applications. In fact, whether a parametric model works well depends on the real situations. For instance, in TOEFL test, 3PL model is preferred since test takers are trained to guess the correct answer if they do not know how to respond. But for some school-level tests, on the other hand, 2PL or Rasch model would be better if test takers have relatively high latent abilities and they do know the correct answers instead of guessing.

2.4 Kernel Smoothing Model

As we mentioned before, parametric models impose restrictive functional form of IRF. However, the observed data may fail to follow these form sometimes. For example, Ramsay[40] shows a figure of estimated ICCs of a test in an introductory psychology course in McGill University. Unfortunately, this graph contains some decreasing curves, which obviously violates the model assumption of monotonicity. In this case, if parametric models are applied, we would definitely obtain ridiculous predictions. Since non-parametric methods require less model assumptions, they are able to fit those data set that the parametric models are unable to do. Meanwhile, they perform well in identifying model misfit, and hence are taken into account for assessing parametric models. There are various non-parametric IRT models such as kernel smoothing model[41] and spline regression[42]. Because spline regression is much more complicated, so we only consider kernel smoothing model, the simpler one, in this thesis. We will present kernel smoothing model in this chapter and execute corresponding kernel smoothing model checking method in the following chapters.

The kernel smoothing model tries to link the probability of response to items correctly with the estimated latent ability through kernel density estimator. The mathematical equation is

$$P(X_i = 1|\theta) = \sum_{j=1}^{n} w_{ij}(\theta) x_{ij}, \tag{2.4}$$

where

$$w_{ij}(\theta) = \frac{K(\frac{\theta-\theta_j}{h_i})}{\sum_{k=1}^{n} K(\frac{\theta-\theta_k}{h_i})}. \tag{2.5}$$

is called *weights*. X_i is binary, namely, $X_i = 1$ denotes responding correctly to the i-th item and $X_i = 0$ means incorrect response to the i-th item. The value of x_{ij} is 1 or 0, indicating whether the response for the j-th test taker to the i-th item is correct or not, respectively. It is seen that the above equation is in fact a weighted average. The $K(t)$ function in w_{ij} is *kernel function*, which is non-negative, continuous and non-increasing when t becomes further from zero. θ_j is the quantile of distribution of latent ability. h_i is the bandwidth of i-th item and it controls how smooth the estimated ICC of i-th item is. If h_i is small, the under-smoothing will be obtained, and in contrast, large h_i will result in over-smoothing. Therefore, the

kernel function, bandwidth and observed data set determine shape and smoothness of the non-parametric estimated ICCs.

Usually there are three kernel functions which are used in $w_{ij}(\theta)$:

- Uniform Kernel Function:

$$K(t) = 0.5, \quad |t| \leq 1, \text{ and 0 otherwise.} \tag{2.6}$$

- Quadratic Kernel Function:

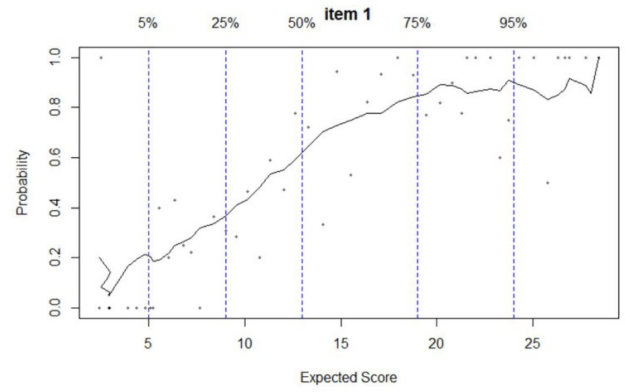
$$K(t) = 0.75(1 - t^2), \quad |t| \leq 1, \text{ and 0 otherwise.} \tag{2.7}$$

- Gaussian(Standard Normal) Kernel Function:

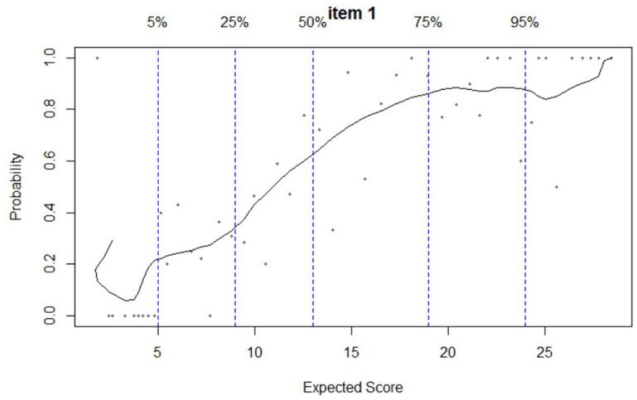
$$K(t) = \exp(-\frac{t^2}{2}). \tag{2.8}$$

We apply these three kernel functions on our simulation, and show the plots of estimated kernel smoothing ICC for item 1 based on the above three different kernel functions in Figure 2.1. From the plots, the shapes of estimated kernel smoothing ICCs are similar, but based on smoothness, it is seen that Gaussian kernel function performs best. Actually, we have similar results for other items, so we finally select Gaussian kernel function to estimate non-parametric ICCs.

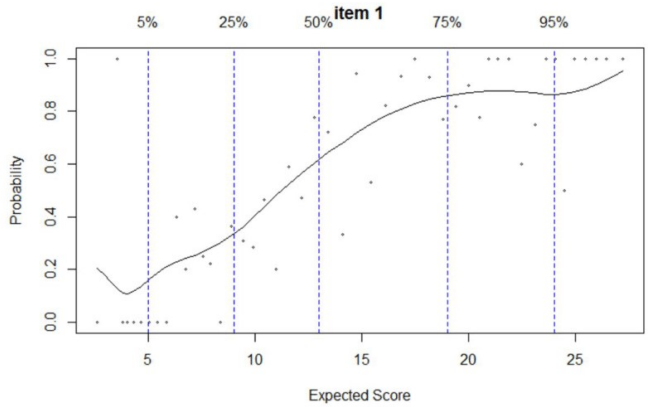
One thing to notice is that the fitted non-parametric ICCs might be decreasing in some cases, which implies the problematic items because it means that low-ability test takers will be more likely to obtain correct response than high-ability test takers, which is ridiculous. Therefore, non-parametric methods can help to identify ill-performed items.



(a) Uniform



(b) Quadratic



(c) Guassian

Figure 2.1: Kernel Smoothing ICC for Item 1 By Different Kernel Functions.

3. Model Checking Methods

In this chapter, we aim to introduce three methods for checking whether the three parametric models could fit the data well through item fit. Item fit means to test the fit of data by investigating information from each item in IRT framework. In fact, there are considerable amount of work concentrating on item fit (e.g., Stone and Zhang[49]). However, most of them just utilize statistical test procedures for goodness-of-fit, and by Hambleton and Han[21], the graphic diagnosis tools are very limited. Due to this fact, our work is to explore graphic diagnosis tools in IRT model assessment. At first, the three checking methodologies that we are going to involve are

1. Observed v.s expected item scores.
2. Randomized quantile residuals.
3. Kernel smoothing method for IRT model.

These methods are going to be applied to our simulation study. Recall that the parametric models must be fitted before we assess them, and two cases for model fitting are considered:

1. In-sample: All the data is used to fit the models.
2. Out-of-sample: The data will be divided into training data set and validation data set randomly. Only the former is for fitting the models.

Actually, the out-of-sample is based on the idea of *simple cross validation*, that is, the raw data is randomly separated into training and validation data set. The former is for training the model and the latter is applied on model checking. More details of in-sample and out-of-sample could be found in chapter 4. Actually, the idea of considering these two cases are from Daniel[12]. In our simulation, for the validation data set, the estimated test takers' latent abilities cannot be gained through model training process, we will estimate them by their sum scores. The basic idea is to rank these sum scores, find the corresponding percentiles and transform percentiles to estimated latent abilities. During this procedure, the latent abilities of validation data set is assessed in the following two ways:

- They are estimated in the entire test takers.
- They are estimated separately.

Notice that the estimated latent abilities in above two situations will be different. This is because the percentile of the same sum score is going to be distinct, which results in the distinct estimation. We point out that in practice, the way of estimating latent traits of validation data set depends on the real situation and we cannot say that estimation based on the first way is better than the second way or vise versa.

When our simulation is formed, we estimate all the parameters for Rasch, 2PL and 3PL models. After that, The expected response pattern matrix for in-sample as well as out-of-sample could be gained respectively. These processes are carried out by *rstan*. Next, the diagnosis plots can be generated. Note that, for in-sample, all the test takers' latent abilities can be estimated by rstan at the same time, while in comparison, only the estimated latent abilities of training data set are received in out-of-sample framework. Before we explain the three model checking methods, a brief description of rstan is given first.

3.1 Rstan

Stan can be called through *rstan* package in R. It is a programming language and provides Bayesian inference by Markov Chain Monte Carlo (MCMC) method such as Hamiltonian Monte Carlo sampling. The main function in rstan package that we will call is **stan()**. To apply this function to the data, several arguments are needed:

1. file: this is a file consisting of stan modelling language.
2. data: it is a list of all input data.
3. iter: the number of iterations for each Markov Chain.
4. chains: the number of Markov Chains.

stan() enables us to get MCMC samplings of all parameters from posterior distributions. By default, the number of MCMC samplings equals to half of iterations because the other half iterations is for warming up and discarded finally. The mean of these samplings are estimated

values of parameters. These estimated parameters are not only involved in computation of expected probability of responding correctly to the given item for each test taker but also generating expected response pattern matrix.

The expected response pattern matrix is the matrix that saves expected scores of each test takers with respect to each item, and the individual expected score is between 0 and 1. in fact, this matrix is generated by rstan and the algorithm is in the following:

1. step 1: calculate estimated success probability for each test taker to each item.

2. step 2: Generate a binary sequence with length of half of total iterations based on Bernoulli distribution with success probability equals to each estimated success probability in step 1.

3. step 3: Take the average of the sequence in step 2, and this average value is the expected score.

In fact, when all the parameters are estimated by rstan, it completes the model fitting process. Basic explanation and examples of applying rstan for fitting IRT models could be found Luo and Jiao[32]. The next step is to execute model checking process. As we mentioned at the beginning, there are three ways and we'll introduce them now.

3.2 Observed v.s Expected Item Sum Scores

Proceeding comparison between observed value and expected values from fitted models serves a very straightforward way in model checking context. In fact, it is an existed methodology which was used many times before, for instance, Swaminathan[53]. In our scenario, since we focus on item fit, the observed as well as expected item sum scores are considered. Here the observed item sum scores are the sum of columns of observed response pattern matrix, and the expected item sum scores are the corresponding values from expected response pattern matrix. The main idea is to check whether the expected item sum scores will be close to observed item scores. We implement this comparison by plotting observed and expected item sum scores in the xOy plane. The former is in the horizontal axis and the latter is put in the vertical axis. If the model fits the data adequately, the pair of observed and expected

item sum scores would almost fall into the diagonal of first quadrant, which we call perfect line. The plot of comparison of our simulation based on this method would be shown in the *Simulation Study* chapter. Unfortunately, it has been shown that such comparison fails to serve as a good way to detect model misfit, and this can be seen in our plot on the next chapter as well. Next, we are going to present a new methodology, that is, *Randomized Quantile Residual (RQR)*.

3.3 Randomized Quantile Residual

Randomized Quantile Residual (RQR) was first proposed by Dunn and Smyth[8]. They actually tried to deal with response variables which are discrete or take on a small number of distinct values because when variables are discrete, the corresponding residuals with respect to observed response values will be parallel in the plot, resulting in difficulty in identifying model misfit. The definition of RQR is given now. We mention here that this definition is based on Feng et al.[9].

Definition Let $Y_1, Y_2, \cdots, Y_n$ be independent identical distributed random variables with realizations $y_1, y_2, \cdots$. Let $F(y; \mu_i, \phi)$ and $p(y; \mu_i, \phi)$ be the CDF and PMF for each Y_i respectively, where $\mu_i = E(Y_i)$ and ϕ is a parameter vector which is common to all Y_i. Let U_i's be independent random variables such that U_i follows uniform distribution on $(0, 1)$ for each i. Define

$$F^*(y_i, U_i; \hat{\mu}_i, \hat{\phi}) = \begin{cases} F(y_i; \hat{\mu}_i, \hat{\phi}), & \text{if } F \text{ is continuous,} \\ F^-(y_i; \hat{\mu}_i, \hat{\phi}) + U_i p(y_i; \hat{\mu}_i, \hat{\phi}), & \text{if } F \text{ is discrete.} \end{cases} \tag{3.1}$$

where $\hat{\mu}_i$ and $\hat{\phi}$ are estimates of μ_i and ϕ, and $F^-(y_i; \hat{\mu}_i, \hat{\phi})$ is the lower limit of F at y_i, namely, $F^-(y_i; \hat{\mu}_i, \hat{\phi}) = \lim_{y \to y_i^-} F(y; \hat{\mu}_i, \hat{\phi})$.

Definition The Randomized Quantile Residual (RQR) $r_{q,i}$ is defined as follows:

$$r_{q,i} = \Phi^{-1}(F^*(Y_i, U_i; \hat{\mu}_i, \hat{\phi})). \tag{3.2}$$

where Φ^{-1} is the quantile function of standard normal distribution (i.e., Φ is the cumulative distribution function of standard normal distribution).

Notice that Φ can be cumulative distribution function of other distribution, for instance, uniform distribution. But we use standard normal distribution in this thesis because RQRs could be obtained easily from *qnorm()* function in R.

An important property of $r_{q,i}$ is that *the $r_{q,i}$ from true model will approximately follow standard normal distribution for each i*. We will prove this fact in the appendix and we mention here that this proof is based on the work of Feng et al.[9].

In practice, checking the normality of $r_{q,i}$ is usually implemented by checking the distribution of sample drawn from $r_{q,i}$. Since drawing samples from $r_{q,i}$ involves drawing sample from U_i, the sample size should be large enough to guarantee that the sample drawn from U_i are approximately uniform distributed so that the sample of $r_{q,i}$ from true model would follow normal distribution approximately.

Checking the normality of RQRs can be achieved by displaying *Q-Q plot* to detect whether there is apparent distinction between sample RQR quantiles and theoretical quantiles (quantiles of standard normal distribution). RQRs from the true model would match the the latter almost perfectly, which means they fall into the perfect line mostly.

3.4 Kernel Smoothing Checking

We introduced kernel smoothing model in the last chapter.The kernel smoothing is a nonparametric technique for describing the relationship between latent traits and probability of obtaining correct response to the given item. Since parametric models rely too much on the parametric form of IRF (e.g., logistic IRF or normal ogive IRF)and several model assumptions, an advantage of kernel smoothing is that it only requires minimum model assumptions, and does not require a specific form of IRF (Meijer et al.[38]), it is just a weighted average. Hence, when the data fails to meet the parametric model assumptions, kernel smoothing model could be considered to fit the data. Also, due to the minimum model assumptions that it is based on, the kernel smoothing model is more accurate in reflecting the relationship of latent traits and probability of responding correctly to the given item than that of parametric models, so it is suggested for assessing parametric model-data fit (Van der Linden[54]). Another practical advantage of kernel smoothing is that there is a

R package called KernSmoothIRT. This package enables us to easily obtain the estimated kernel smoothing ICCs which can be utilized for parametric model checking. In this thesis, we are going to use this package to estimate kernel smoothing ICCs in light of our simulation, then these ICCs are applied for parametric model checking. We point out here that there have been several R packages which are able to estimate parametric IRT ICCs, for instance, Chalmers explains *mirt* package[6], and Rusch introduces other packages such as *eRm* and *ltm*[44], but they are not considered in this thesis. First, we introduce the *KernSmoothIRT* Package and its main function *ksIRT()*.

KernSmoothIRT is a specific R package for estimating nonparametric ICCs of IRT models, which is developed on the basis of Ramsay's **TestGraf**, a program of graphical analysis of multiple choice test. The main function in this package that we are going to use is *ksIRT()*.

Applying *ksIRT()* requires three principal arguments: response pattern matrix, correct answer response (key) for each item and type (format) of items. In our simulation, both key and format equal to 1. It also needs type of kernel function as well as bandwidth. As we mentioned previously, The Gaussian kernel function is used here, which is actually the default type by *ksIRT()*. With respect to the bandwidth, in fact, if we consider Gaussian kernel function, there are two kind of bandwidth: *Silverman* and *CV*. *Silverman* means Silver rule of thumb and it is the default bandwidth because it serves as the optimal bandwidth of Gaussian kernel and its calculation formula is

$$h_{opt} = 1.06\sigma_\theta n^{-\frac{1}{5}}. \tag{3.3}$$

where σ_θ is the sample standard deviation and n is the sample size. If we select *Silverman*, all the items have the same bandwidth. The other selection *CV* refers to cross validation bandwidth. It is calculated by applying leave-one-out algorithm and minimizing the cross-validation statistic for individual item, resulting in different bandwidth for distinctive item. The details could be found in Angelo Mazza[35]. We tried these two bandwidth to our simulation and it seems *Silverman* works better, so it is selected finally.

The last point here is the range of latent ability. For parametric models, latent abilities are assumed to follow standard normal distribution. In the kernel smoothing scenario, however, latent traits are measured by expected sum scores, which indicates that each assessment is

between 0 and full points of the test.

The strategy for utilizing nonparametric checking technique is to contrast non-parametric ICCs to estimated success probabilities of test takers regarding each item from parametric models via graphs. Noticeable difference suggests the misfit of parametric ones.

The graph of nonparametric ICCs is acquired directly by plotting *ksIRT*. The estimated success probabilities of three parametric models are followed from their IRF respectively and parameters in each IRF are estimated by rstan. Recall that in the kernel smoothing context, since latent traits are assessed by expected sum scores, expected sum score for test takers from parametric models are needed at the same time. As a matter of fact, this can be earned by adding up each row of expected response pattern matrices.

The checking steps are below:

1. Calculate estimated probabilities of three candidate parametric models.
2. Fit the nonparametric ICC through the kernel smoothing.
3. Plot the fitted curve that obtained in step 2.
4. Plot the values gained in step 1 on the same graph.
5. Do comparison between nonparametric and parametric models.

Note that the x-coordinate of plot in step 3 above represents the expected sum scores and y-coordinate is the expected success probabilities from kernel smoothing with respect to the given item.

4. Simulation Study

In this chapter, we will apply all the checking methods in chapter 3 to our simulated data, a 526×30 simulated response pattern matrix with binary entries. The simulated matrix is created on the basis of 2PL model first. Then we fit our parametric models, and each checking method is proceeded thereafter. Next, the result of each checking method will be shown via plots. Finally we will give conclusion.

4.1 Data Generation

We use the *Math Placement Test (MPT)* in the University of Saskatchewan, which is a qualification test aiming to examine whether test takers' mathematical background is enough for learning advanced college-level math courses, as our reference to generate the simulation. MPT comprises 30 multiple choice questions and each question is binary, i.e, the response is either correct (1 point) or incorrect (0 point) and there's no partial point. The full score is 30. Finally, both test taker's response patterns and their sum scores are received and evaluated by committees for qualification. In MPT, since students prefer not to respond if they totally do not know how to figure out the question, so we select 2PL model instead of 3PL model to generate simulation. There were 526 students who took part in MPT in the University of Saskatchewan last year, so in our simulation, we select 526 as the number of test takers and 30 as the number of items. So our simulation will be a 526×30 binary matrix, the row of this matrix represents test takers and column is items. We emphasize here that in our simulation, the sample size for test takers, which is 526, is large enough to fit parametric IRT models, however, in the general case, this sample size might not be sufficient large (e.g. GRE test or TOEFL test). To generate this simulated response pattern matrix, true values of model parameters as well as latent abilities are needed first. Notice that item discrimination α follows log-normal distribution, and if $\alpha > 1.7$, the ability of the item to distinguish latent abilities is considered to be very high[3], hence, all the item discrimination are generated through log-normal distribution with mean 0 and standard deviation 0.3 because it can

prevent α's from being too large. Second, both item difficulty β and latent ability θ follow standard normal distribution, which means they are generated from $N(0,1)$. When this step is completed, the simulated binary response pattern matrix is able to be created through 2PL model.

We list the steps of generating simulation here:

1. Generate item discrimination α_j, item difficulties β_j and latent abilities θ_i, $i = 1, ..., 526$, $j = 1, ..., 30$, from log-normal, and standard normal distribution respectively, that is,

$$\alpha_j \sim \text{log-normal}(0, 0.3), \quad \beta_j \sim N(0,1), \quad \theta_i \sim N(0,1). \tag{4.1}$$

2. Calculate p_{ij}, the probability for i-th test taker to respond correctly to j-th item, by 2PL model.

3. For each p_{ij}, generate a response O_{ij} from binomial distribution

$$O_{ij} \sim \text{Binomial}(p_{ij}). \tag{4.2}$$

   and save them in a 526×30 matrix. The rows represent test takers and columns the items.

The matrix in step 3 is our simulated response pattern matrix, which is binary. Hence each Q_{ij} is 0 or 1.

Since we will consider in-sample and out-of-sample cases and only give a brief description about them in chapter 3, now we provide more details.

Usually, the observed data set will be randomly separated into training data set, which is used to fit/estimate/train the models, and validation data set being utilized to evaluate the fitted model (Sometimes training data set and validation data set could be the same). The sample size of training data set should be large enough so the model could be estimated as accurate as possible.

4.1.1 In-sample v.s Out-of-sample

1. *In-sample case.* For in-sample case, the training data set is as same as validation data set. When the estimated model is obtained based on training data set, the same data

set is used again to evaluate the model-data fit, that is, the training data set is applied twice, one for fitting the model, one for evaluating the fitted model. The result is that the difference between training data set and model predictions will be small. So it is hard to detect model misfit in this case.

2. *Out-of-sample case.* For out-of-sample case, the training data set will be different from validation data set. Only the training data set is applied for model fitting, and validation data set is considered as a new data set to assess the fitted model. Generally, the model predictions in out-of-sample case will be different from the observed data if the model fails to fit the data well. Out-of-sample is usually utilized for assessing model-data fit.

In our simulation study, the observed data set is a binary response pattern matrix, with row the test takers and column the items, so it is a two-dimensional data set (the first dimension is test taker and the second dimension is item). We are going to sample test takers randomly and use these sampled test takers' response pattern as the training data set to fit parametric IRT models. In the in-sample situation, we take entire test takers as our sample, so the whole binary response pattern matrix is the training as well as validation data set. In the out-of-sample situation, we will randomly select 400 test takers as our sample, and the training data set is the response patterns of these 400 test takers. The rest 126×30 matrix will be the validation data set. Note that the training data set is different for in-sample and out-of-sample case in this thesis.

4.2 Results of Simulation Study

4.2.1 Observed v.s Expected Item sum Scores Result

As we mentioned before, item sum scores are sum of each column of observed or expected response pattern matrix. The number of item sum scores equals to the number of items. The plot of observed item sum score v.s expected item sum score of three parametric models is displayed in Figure 4.1.

The Figure 4.1 contains three plots for in-sample and out-of-sample case. The horizontal

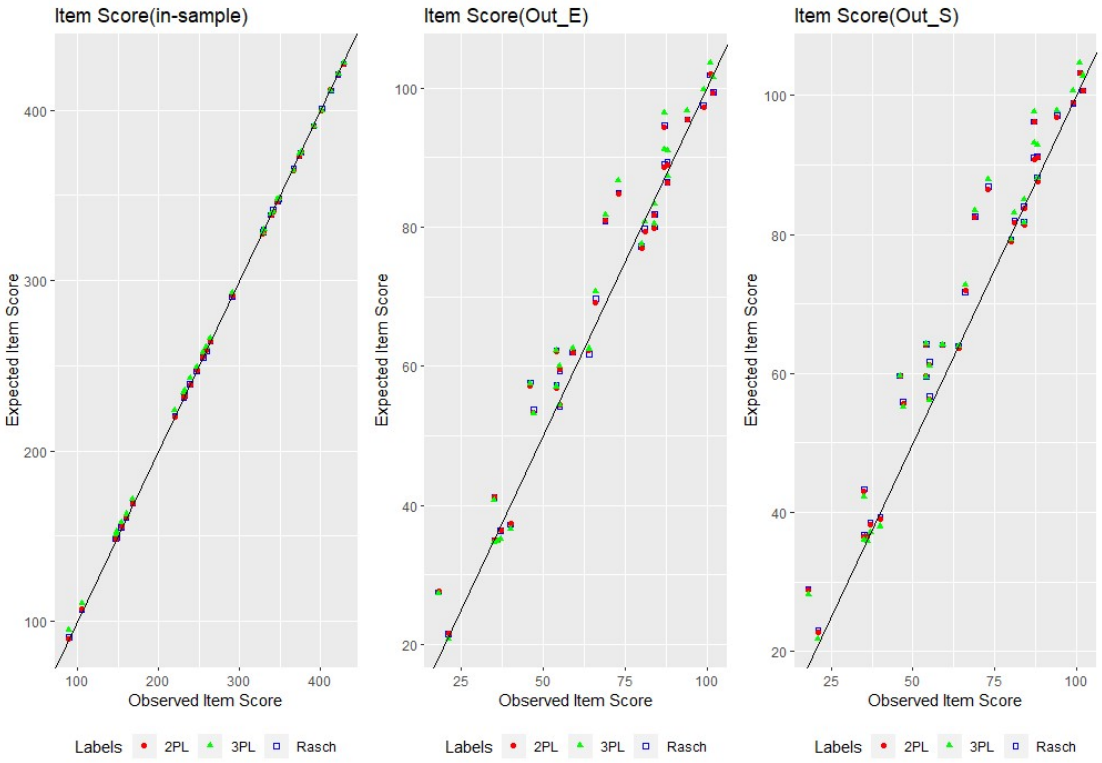


Figure 4.1: Observed v.s Expected Item Sum Scores.

line represents observed item sum score and the vertical line is expected item sum score. The notation 'Out_E' means out-of-sample with the latent abilities of validation set estimated in the entire test takers and 'Out_S' indicates the same thing but estimated separately (in the validation set itself).

The plot of in-sample case (the very left one) implies that the expected values from all three models almost match the observed values perfectly, except there's a little bit deviation for 3PL model at the lower tail part. While for out-of-sample case, we do see departures for all parametric models and it seems that there's no obvious differences between these departures. So it is difficult for us to detect model misfit through visualization by means of this method. As a matter of fact, this method was not effective for goodness-of-fit, though it is quite straightforward. The reason for this is that sum scores collapses individual information regarding each item.

4.2.2 Randomized Quantile Residual(RQRs) Result

1. *In-Sample Case*. In this situation, RQRs is obtained on the basis of the entire simulated response pattern matrix (526×30). Rstan is able to provide samplings of each parameters via their posterior distribution. We are going to make use of all the samples of each model parameter and posterior means of all test takers' latent abilities to generate RQRs for each item. The reason that we consider posterior means of each θ is that involving all samples of θ results in huge number of RQRs, which both consumes lots of storage and takes long time for plotting. Finally these RQRs are plotted to detect model misfit.

In rstan, for each model, we use 2 Markov chains with each chain 2000 iterations. The total iterations is $2 \times 2000 = 4000$ and number of samplings for each model parameter equals to 2000. The number of estimated latent abilities is 526 actually. The estimated success probabilities $\hat{p}_{ijk}$ for i-th test taker with respect to k-th sampling of j-th item are calculated via three models

- Rasch Model

$$\hat{p}_{ijk}^{\text{Rasch}} = \frac{\exp(\hat{\theta}_i - \hat{\beta}_{jk})}{1 + \exp(\hat{\theta}_i - \hat{\beta}_{jk})}. \tag{4.3}$$

- 2PL Model

$$\hat{p}_{ijk}^{\text{2PL}} = \frac{\exp(\hat{\alpha}_{jk}(\hat{\theta}_i - \hat{\beta}_{jk}))}{1 + \exp(\hat{\alpha}_{jk}(\hat{\theta}_i - \hat{\beta}_{jk}))}. \tag{4.4}$$

- 3PL Model

$$\hat{p}_{ijk}^{\text{3PL}} = \hat{\gamma}_{jk} + (1 - \hat{\gamma}_{jk}) \frac{\exp(\hat{\alpha}_{jk}(\hat{\theta}_i - \hat{\beta}_{jk}))}{1 + \exp(\hat{\alpha}_{jk}(\hat{\theta}_i - \hat{\beta}_{jk}))}. \tag{4.5}$$

$\hat{p}_{ijk}$'s are stored in 3-dimension arrays and we have three such arrays in total(one for each model).

The next step is to produce RQRs. The rule is

$$\text{RQR} = \begin{cases} \Phi^{-1}(u_{ijk}), & u_{ijk} \in \text{Uniform}(0, 1 - \hat{p}_{ijk}] \text{ if } O_{ij} = 0, \\ \Phi^{-1}(u_{ijk}), & u_{ijk} \in \text{Uniform}(1 - \hat{p}_{ijk}, 1] \text{ if } O_{ij} = 1. \end{cases} \tag{4.6}$$

where Φ^{-1} is the inverse cumulative distribution function of Gaussian distribution, u_{ijk} is a random number of uniform distribution on the interval $(0, 1-\hat{p}_{ijk}]$ or $(1-\hat{p}_{ijk}, 1]$. The RQRs are collected in a 3-dimension array as well, with size $526 \times 2000 \times 30$, indicating test takers, samplings and items respectively. For each item, the corresponding RQRs is a 526×2000 matrix, for example, the RQRs of i-th item is

$$\begin{matrix} RQR^i_{1,1} & RQR^i_{1,2} & ... & RQR^i_{1,2000} \\ RQR^i_{2,1} & RQR^i_{2,2} & ... & RQR^i_{2,2000} \\ ... & ... & ... & ... \\ RQR^i_{526,1} & RQR^i_{526,2} & ... & RQR^i_{526,2000} \end{matrix} \tag{4.7}$$

each of these RQRs matrices is treated as a vector when we plot them, entailing the RQRs number is $526 * 2000 = 1,052,000$ per item.

2. *Out-of-Sample Case*. Different from in-sample case, we fit our model by training data set, and this time, only samplings of model parameters are needed. Since RQRs creation involves latent abilities of validation set, it requires estimated values of them and this results from transformation of rank of sum scores.

The formation of RQRs follows the same rule as in-sample case except the size. We just have 126 test takers in validation set, RQRs for a single item is a 126×2000 matrix. Because latent abilities are evaluated in two ways, each item will have two corresponding RQRs.

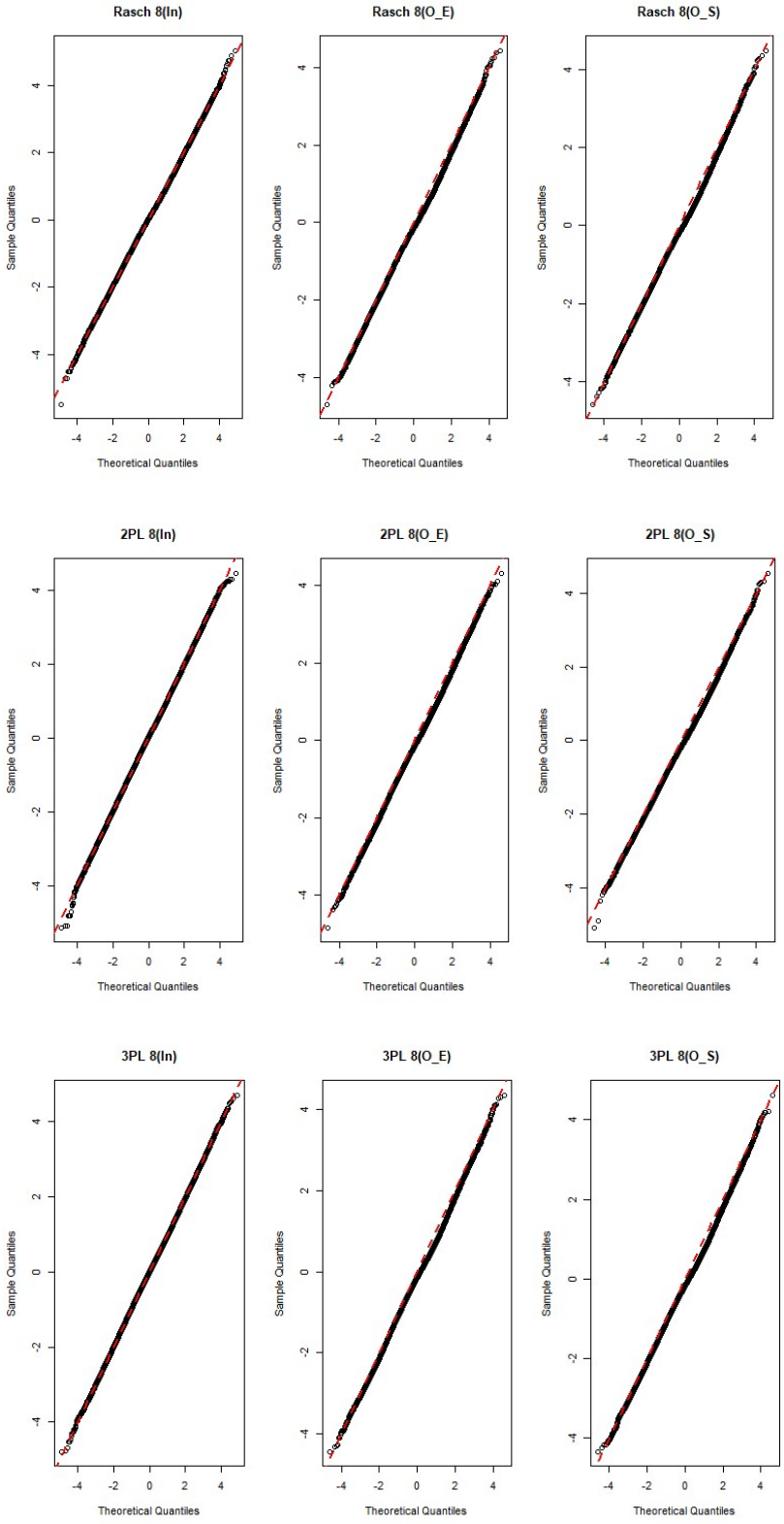


Figure 4.2: RQR Checking Plot for Item 8.

Figure 4.2 exhibits RQRs of item 8 of three models for all cases.

Similar to Figure 4.1, the very left plot stands for in-sample and the other two are out-of-sample. It is seen from above figures that RQR plots of in-sample for all three models have no obvious deviation. As a matter of fact, this is true for other items as well. The reason is that we train our optional models by the usage of entire simulated response pattern matrix, and use these information again for creating RQRs. Hence, the same data is actually utilized twice, resulting in hardness in identifying model misfit.

When it comes to out-of-sample, we do observe apparent departure in Rasch and 3PL model from Figure 4.2. The analogous deviation also happens in item 10, 14, 18, 21, 23, 25, 26 for these two wrong models. However, Figure 4.2 shows deviation of RQRs for 2PL model at the same time, and it is likewise for some other items (all the other RQR plots are shown in the appendix C).

These outcomes implies out-of-sample outperforms than in-sample case, in that the latter fails to detect any wrong model at all, while the former, in contrast, as indicated in the plot, at least discovers model misfit through several items.

4.2.3 Kernel Smoothing Checking Result

By the idea as well as checking steps of this methodology in chapter 3, expected person sum scores and corresponding estimated success probabilities from parametric models would be calculated. Each person expected sum score is a decimal between 0 and 30. This is different from observed ones, which are integers with same range. In the context of kernel smoothing assessment, success probabilities is estimated on the basis of posterior mean of model parameters.

Plots of the kernel smoothing for item 29 are shown in Figure 4.3. Expected person scores are placed in the x-axis and predicted success probabilities the y-axis. the black curve serves as the fitted kernel smoothing item characteristic curve. We pair person scores and probabilities then place them on the same graph for comparison. Here purple, red and green points denote Rasch, 2PL and 3PL respectively. It is clear that Rasch model demonstrates noticeable departure. 2PL and 3PL basically follow this fitted non-parametric curve, but the latter indicates more deviation than the former when expected person scores are below 12.

We found resembling consequences in most other items, that is, Rasch is the worst model in fitting the data, and 2PL performs best. The plots of kernel smoothing checking for all other items are in the appendix D.

There are still some black dots near the non-parametric curve, which are called *grouped subject scores*[35]. Angelo et al.[35] gave detailed explanation about how to calculate them and we simply state the idea here. The range of latent abilities θ is split into a finite grid of q values $\theta_1, ..., \theta_q$ with equal distance δ first. Afterwards, all test takers are grouped by observed sum scores and these sum scores will be ranked and transformed to estimated latent abilities, denoted by $\hat{\theta}_j, \quad j = 1, ...n$, where n is the number of test takers. Furthermore, two sequences of q values are defined

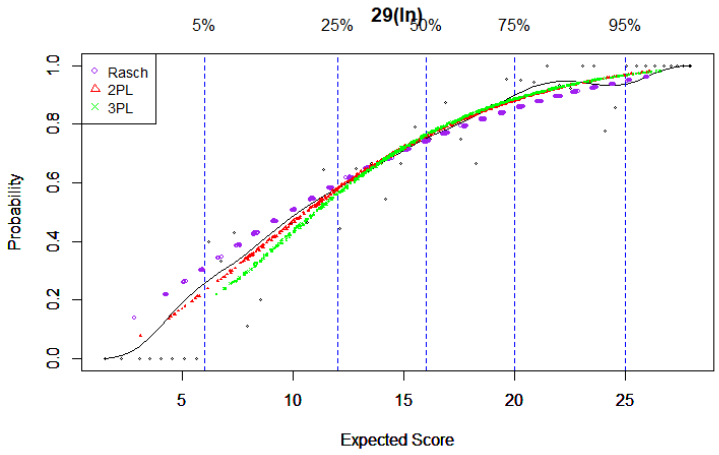
$$\widetilde{y}_{si} = \sum_{j=1}^{n} I_{[\theta_s - \frac{\delta}{2}, \theta_s + \frac{\delta}{2}]}(\hat{\theta}_j) y_{ij}, \quad v_s = \sum_{j=1}^{n} I_{[\theta_s - \frac{\delta}{2}, \theta_s + \frac{\delta}{2}]}(\hat{\theta}_j). \tag{4.8}$$

where i is the i-th item, $s \in \{1, 2, ..., q\}$ and I is the indicator function. The estimated probability of responding correctly with respect to i-th item is

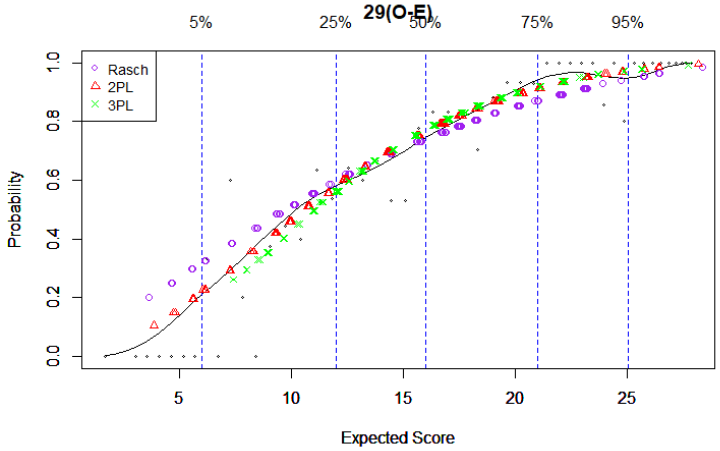
$$\hat{p}_i(\theta) \approx \frac{\sum_{s=1}^{q} K(\frac{\theta - \theta_s}{h_i}) \widetilde{y}_{si}}{\sum_{s=1}^{q} K(\frac{\theta - \theta_s}{h_i}) v_s}, \quad \theta \in \{\theta_1, ..., \theta_q\}. \tag{4.9}$$

These $\hat{p}_i$ are grouped subject scores.

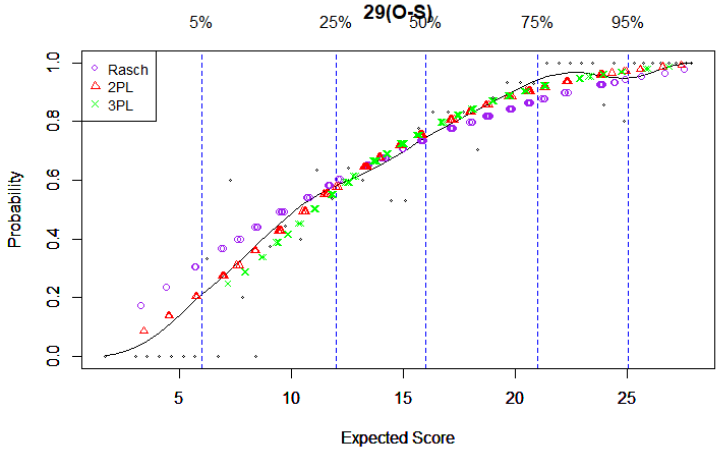
Therefore, it seems that, owing to what we observe in the plots, non-parametric serves as an effective way in identifying model misfit.



(a) Item 29: In-sample



(b) Item 29: Out-of-sample-E



(c) Item 29: Out-of-sample-S

Figure 4.3: Kernel Smoothing Checking Plot for Item 29.

5. Conclusion and Future Work

In this thesis, we apply three methodologies, i.e, observed v.s expected item sum scores, randomized quantile residual (RQR) as well as kernel smoothing checking (the nonparametric checking method) to identify model misfit. The data is a binary simulated response pattern matrix from two-logistic (2PL) model and our parametric model candidates are Rasch, 2PL and 3PL model. The in-sample and out-of-sample situations are considered in the model assessment process as well. Our preliminary results of simulation study elucidates that comparison between observed and expected item sum scores fails to be effective in performing model checking, while for RQRs, out-of-sample outperforms in-sample case to detect model misfit, and it seems that the nonparametric way, kernel smoothing method, serves as a compelling way to be employed in assessing parametric models.

Our future work is going to concentrate on the following aspects:

1. We will implement these checking methods on some real data. As what we referred to previously, the only number of test takers and items of our simulation is from some real test. We will use real test data later to fit parametric models and carry out model checking, especially the nonparametric method. This can help both the exam designers and committees to check whether there are problematic items and assess test takers' latent traits so they can have a better understanding of the entire data features.

2. We just focus on item fit in this simulation. However, there have been different kinds of other model fit, for instance, person fit and overall fit. One of our future goals is to involve these fit in our model checking context.

3. All the parametric models here only concerns unidimensional latent space. However, in some cases, it is more realistic for us to assume that a person's responses should be described by his or her multiple latent abilities. For example, a non-native speaker's performance in the GRE math test are not only related to the math ability but also reading capability. Thus, we are going to explore model assessment for multi-dimensional parametric model in our future work.

4. Finally, we assume the latent abilities follows standard normal distribution. Unfortunately, in real situations, it's not the case. For instance, Ramsay shows the empirical distribution of sum scores for a standard exam which was skewed[42]. Due to this reason, he proposes a new concept called *optimal score*[42], being based on *the tilted scaled β distribution*[40]. By Ramsay, the optimal score is a more accurate estimation of latent abilities than sum score. We will consider optimal score instead of sum scores in the model assessment of our next work, and we believe this can enable us to improve sensitivity of our checking methods in detecting model misfit.

References

[1] R.J.de Ayala. *The Theory and Practice of Item Response Theory*. The Guilford Press, 2009.

[2] Wei Bai. Randomized quantile residual for assessing generalized linear mixed models with application to zero-inflated microbiome data. Master's thesis, University of Saskatchewan, 2018.

[3] Frank B. Baker. *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, 2001.

[4] R.D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37:29–51, 1972.

[5] Bryonna Bowen. Goodness of fit via residual plots in item response theory. Master's thesis, University of South Carolina, 2018.

[6] R.Philip Chalmers. mirt, a multidimensional item response theory package for the r environment. *Journal of Statistical Software*, 48(6), 2012.

[7] Jeffrey Douglas and Allan Cohen. Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, 25(3):234–243, 2001.

[8] Peter K. Dunn and Gordon K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

[9] Cindy Feng, Alireza Sadeghpour, and Longhai Li. Randomized predictive p-values: A versatile model diagnostic tool with unified reference distribution. *https://arxiv.org/abs/1708.08527*, 2019.

[10] Pere J. Ferrando. An irt modeling approach for assessing item and person discrimination in binary personality responses. *Applied Psychological Measurement*, 40(3):218–232, 2016.

[11] J. Fu, D. M. Bolt, and Y. Li. Evaluating item fit for a polytomous fusion model using posterior predictive checks. *Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada*, 2005.

[12] Daniel C. Furr. *Bayesian and frequentist Cross-validation Methods for Explanation Item Response Models*. PhD thesis, University of California, Berkeley, 2017.

[13] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. CRC Press, 2014.

[14] Glas and Dagohoy. A bayesian approach to person irt analysis in item response theory models. *Applied Psychological Measurement*, 27(3):217–233, 2003.

[15] C.A.W. Glas. Detection of differential item functioning using lagrange multiplier tests. *Statistica Sinica*, 8(1):647–667, 1998.

[16] C.A.W. Glas. Modification indices for the 2-pl and the nominal response model. *Psychometrika*, 64:273–294, 1999.

[17] C.A.W. Glas and R. R. Meijer. A person fit test for irt models for polytomous items. *Psychometrika*, 72:159–180, 2007.

[18] C.A.W. Glas and J.C.S. Suarez Falcon. A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27(2):87–106, 2003.

[19] I. Guttman. The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society*, 29:83–100, 1976.

[20] L. Guttman. The cornell technique for scale and intensity analysis. *Educational and Psychological Measurement*, 7:247–280, 1947.

[21] R. Hambleton and N. Han. Assessing the fit of irt models: Some approaches and graphical displays. *In annual meeting of the National Council on Measurement in Education, San Diego, CA*, 2004.

[22] R.K. Hambleton and H. Swaminathan. *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff, 1965.

[23] R.K. Hambleton and H.J. Swaminathan, H.and Rogers. *Fundamentals of Item Response Theory*. Sage, 1991.

[24] Ronald K. Hambleton and Russell W. Jones. Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3):38–47, 1993.

[25] Steven T. Kalinowski. A graphical method for displaying the model fit of item response theory trace lines. *Educational and Psychological Measurement*, 79(6):1064–1074, 2019.

[26] T. Kang and T. T. Chen. Performance of the generalized $s - x^2$ fit index for polytomous irt models. *Journal of Educational Measurement*, 45:391–406, 2008.

[27] M. Kuhfeld. A posterior predictive model checking method assuming posterior normality for item response theory. *Appl Psychol Meas*, 43(2):125–142, 2019.

[28] Hee Chon Kyong, Won-Chan Lee, and Stephen B. Dunbar. Comparison of item fit statistics for mix irt models. *Journal of Educational Measurement*, 47(3):318–338, 2010.

[29] R. Levy. Posterior predictive model checking for multidimensionality in item response theory and bayesian networks (unpublished). 2006.

[30] T. Li, C. Xie, and H. Jiao. Assessing fit of alternative unidimensional polytomous irt models using posterior predictive model checking. *Psychol Methods*, 22(2):397–408, 2017.

[31] Tie Liang. *An Assessment of The Nonparametric Approach for Evaluating The Fit of Item Response Models*. PhD thesis, University of Massachusetts Amherst, 2010.

[32] Yong Luo and Hong Jiao. Using the stan program for bayesian item response theory. *Educational and Psychological Measurement*, 78(3), 2017.

[33] A. Maydeu-Olivares and H. Joe. Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, 100:1009–1020, 2005.

[34] Alberto Maydeu-Olivares. Goodness-of-fit assessment of item response theory models. *Measurement*, 11:71–101, 2013.

[35] Angelo Mazza, Antonio Punzo, and Brian McGuire. Kernsmoothirt: An r package for kernel smoothing in item response theory. *Journal of Statistical Software*, 58(6), 2014.

[36] Roderick P. McDonald. Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, 6(4):379–396, 1982.

[37] R. McKinley and C. Mills. A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9:49–57, 1985.

[38] Rob R. Meijer and Joost J. Baneke. Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods*, 9(3):354–368, 2004.

[39] M. Orlando and D. Thissen. Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1):50–64, 2000.

[40] James O. Ramsay and Marie Weberg. A strategy for replacing sum scores. *Journal of education and behavioral statistics*, 42(3):282–307, 2017.

[41] J.O. Ramsay. Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56(4):611–630, 1991.

[42] J.O. Ramsay, Juan Li, and Marie Weiberg. Optimal scores: An alternative to parametric item response and sum scores. *Psychometrika*, 84(1):310–322, 2019.

[43] D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–1172, 1984.

[44] Thomas Rusch, Patrick Mair, and Reinhold Hatzinger. Psychometric with r, a review of cran packages for item response theory. 2013.

[45] Alireza Sadeghpour. Randomized quantile residuals for diagnosis of non-normal regression models. Master's thesis, University of Saskatchewan, 2016.

[46] S.Von Schrader, T.N. Ansley, and S. Kim. Examination of item fit indices for polytomous item response models. *Paper presented at the meeting of the National Council on Measurement in Education*, 2004.

[47] S Sinharay. Assessing fit of unidimensional item response theory models using a bayesian approach. *Journal of Educational Measurement*, 42(4):375–394, 2005.

[48] C.A. Stone. Monte carlo based null distribution for an alternative goodness-of-fit test statistic in irt models. *Journal of Educational Measurement*, 37:58–75, 2000.

[49] C.A. Stone and B. Zhang. Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40(4):331–352, 2003.

[50] S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen (Eds). The basis for scalogram analysis. *Measurement and Prediction: Studies in Social Psychology in World War II*, 4, 1950a.

[51] S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen (Eds). Relation of scalogram analysis to other techniques. *Measurement and Prediction: Studies in Social Psychology in World War II*, 4, 1950b.

[52] W. F. Stout. A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, 52:589–617, 1987.

[53] Hariharan Swaminathan, Ronald K. Hambleton, and H.Jane Rogers. Assessing the fit of item response theory models. *Handbook of Statistics*, 26:683–719, 2007.

[54] Wim J. Van der Linden and Ronald K. Hambleton. *Handbook of Modern Item Response Theory*. Springer, 1996.

[55] W. M. Yen. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8:125–145, 1984.

[56] W.M. Yen. Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 8(2):245–262, 1981.

[57] W.M. Yen. Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30:187–213, 1993.

[58] Xiaowen Zhu. *Assessing Fit of Item Response Models for Performance Assessments Using Bayesian Analysis*. PhD thesis, University of Pittsburgh, 2009.

Appendix A

Proof of Inverse CDF Theorem

The *Inverse CDF Theorem* is stated and proved in the following.

Inverse CDF Theorem Let X be a continuous random variable with CDF F_X. Let $U \sim$ Uniform(0,1). Then the random variable Y defined by

$$Y = F_X^{-1}(U) \tag{A.1}$$

is a continuous random variable with CDF F_X, where F_X^{-1} is the inverse function of F_X.

Proof. To show the CDF of Y is F_X, we need to show that,

$$P(Y \leq y) = F_X(y) \tag{A.2}$$

In fact,

$$P(Y \leq y) = P(F_X^{-1}(U) \leq y) = P(U \leq F_X(y)) = F_X(y). \tag{A.3}$$

Appendix B

Proof of Normality of RQRs of True Model

Theorem Let $R_{q,i}$ be the RQR from the true model with true value of μ_i and ϕ, namely,

$$R_{q,i} = \Phi^{-1}(F^*(Y_i, U_i; \mu_i, \phi)) \tag{B.1}$$

Then $R_{q,i}$ will follow exactly standard normal distribution for each i.

Proof. In the proof, We will omit μ_i and ϕ for convenience(i.e., we will write $F^*(Y_i, U_i; \mu_i, \phi)$ as $F^*(Y_i, U_i)$). The proof involves the *Inverse CDF Theorem*. The details of this theorem is given in the appendix A. By the inverse CDF theorem, we need to show $F^*(Y_i, U_i)$ follows uniform distribution on $(0, 1)$ for each i. This is equivalent to show:

if E is an interval contained in $(0, 1)$ and let $m(E)$ be the length of E, then the probability

$$p(F^*(Y_i, U_i) \in E) = m(E) \tag{B.2}$$

Let $E \subset [0, 1]$ be an interval with length $m(E)$.

(i) If F is continuous. Notice that $F^*(Y_i, U_i) = F(Y_i)$ and $F(Y_i) \in (0, 1)$, the probability $p(F(Y_i) \in E)$ is actually the measure of the event of $F(Y_i) \in E \subset (0, 1)$, so it is obvious that

$$p(F^*(Y_i, U_i) \in E) = p(F(Y_i) \in E) = m(E) \tag{B.3}$$

(ii) If F is discrete. Suppose all possible realizations of Y_i are $y_1, y_2...$, then for each $k \in \mathbb{N}$,

$$F^*(y_k, U_i) = F^-(y_k) + U_i p(y_k) \tag{B.4}$$

is a uniform random variable on $(F^-(y_k), F(y_k))$. Notice that

$$\bigcup_{k=1}^{\infty}(F^-(y_k), F(y_k)) = (0, 1) \setminus \bigcup_{k=1}^{\infty}\{F(y_k)\} \tag{B.5}$$

hence

$$(0, 1) = \bigcup_{k=1}^{\infty}(F^-(y_k), F(y_k)) \cup \{F(y_1)\} \cup \{F(y_2)\} \cup ... \tag{B.6}$$

since $(F^-(y_{k_1}), F(y_{k_1})) \cap (F^-(y_{k_2}), F(y_{k_2})) = \emptyset$ if $k_1 \neq k_2$, so

$$\begin{aligned} E &= E \cap (0, 1) \\ &= E \cap (\bigcup_{k=1}^{\infty}(F^-(y_k), F(y_k)) \cup \{F(y_1)\} \cup \{F(y_2)\} \cup ...) \\ &= \bigcup_{k=1}^{\infty}(E \cap (F^-(y_k), F(y_k))) \cup \{E \cap F(y_1)\} \cup \{E \cap F(y_2)\} \cup ... \end{aligned} \tag{B.7}$$

therefore we have

$$
\begin{aligned}
p(F^*(Y_i, U_i) \in E) &= \sum_{k=1}^{\infty} p(F^*(y_k, U_i) \in E) \\
&= \sum_{k=1}^{\infty} p(F^*(y_k, U_i) \in E \cap (F^-(y_k), F(y_k))) + \sum_{k=1}^{\infty} p(F^*(Y_i, U_i) \in \{E \cap F(y_k)\}) \\
&= \sum_{k=1}^{\infty} p(Y_i = y_k) \cdot p(F^*(Y_i, U_i) \in E | Y_i = y_k) + \sum_{k=1}^{\infty} p(F^*(Y_i, U_i) \in \{E \cap F(y_k)\}) \\
&= \sum_{k=1}^{\infty} p(Y_i = y_k) \cdot \frac{m(E \cap (F^-(y_k), F(y_k)))}{p(Y_i = y_k)} + 0 \\
&= \sum_{k=1}^{\infty} m(E \cap (F^-(y_k), F(y_k))) \\
&= m(E).
\end{aligned}
\tag{B.8}
$$

Appendix C

Randomized Quantile Residual Plots

The Appendix C displays RQR checking plots for all items except item 8, so we have totally 29 RQR checking plots. These plots are listed from next page and they will help us to have a better understanding of the conclusion that we obtained in this thesis. The RQR checking plot for each item consists of 9 small plots, representing three parametric models with respect to three different cases. The 'In' in the plot means in-sample case, while 'O_E' and 'O_S' indicates out-of-sample case with latent abilities of validation data set estimated in the 526 test takers and 126 test takers respectively. For in-sample case, the sample size is $526 * 2000 = 1,052,000$ per item, and the sample size is $126 * 2000 = 252,000$ per item for out-of-sample. Finally, we point out that the generation of RQRs comes from the samples of model parameters, since we use MCMC to draw samples, it is necessary for us to check the convergence of Markov chain. If the Markov chain fails to converge, the samples which we draw are unreliable. In our simulation, all the Markov chains are convergent.

It can be seen, from these plots, that when it comes to in-sample case, the RQR fails to assess model-data fit for all items because there is no deviation at all for all parametric models. The reason is that we use the same data set to fit and check the model, meaning that the same data is used twice, so it is hard for us to identify model misfit. As for out-of-sample case, deviations do happen in the RQR checking plots of wrong models (Rasch and 3PL models) through some items (for example, Figure C.9, Figure C.13 and Figure C.20), however, such deviations are also found in the RQR plots of true model (2PL model) through these items and it seems that there is no obvious difference between these deviations for all three parametric models. so we can only conclude that out-of-sample performs better than the in-sample case in assessing model-data fit.

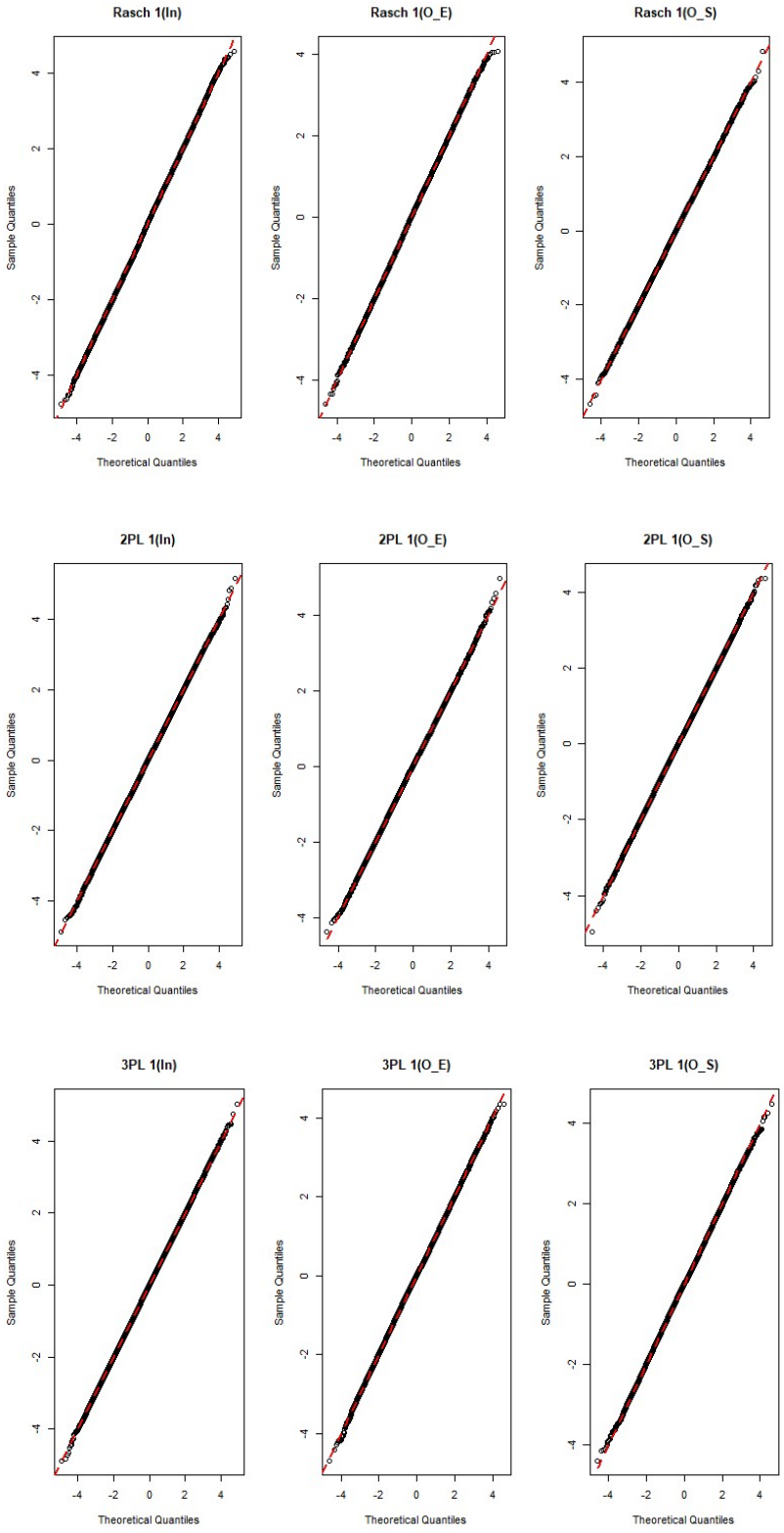


Figure C.1: RQR Checking Plot for Item 1.

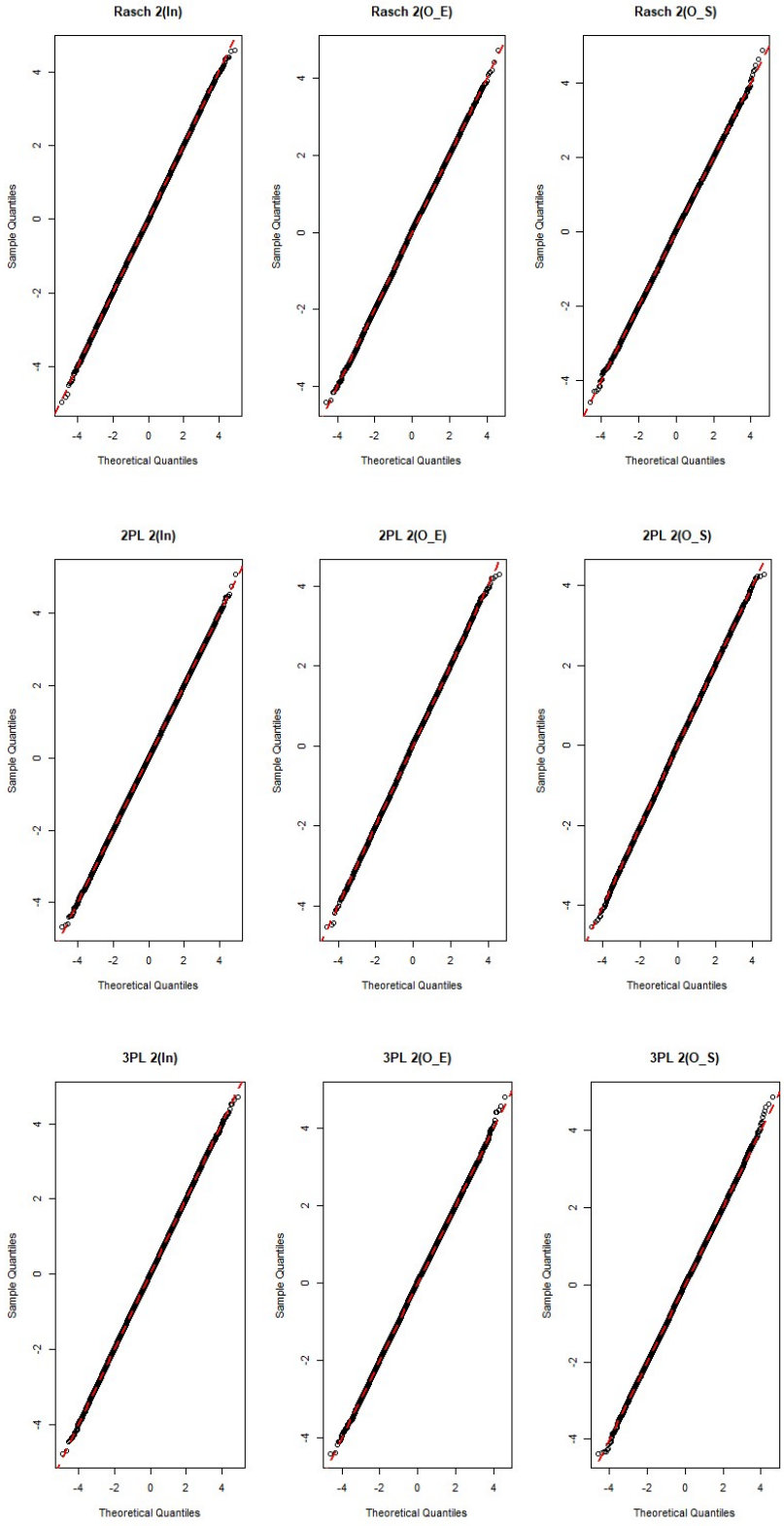


Figure C.2: RQR Checking Plot for Item 2.

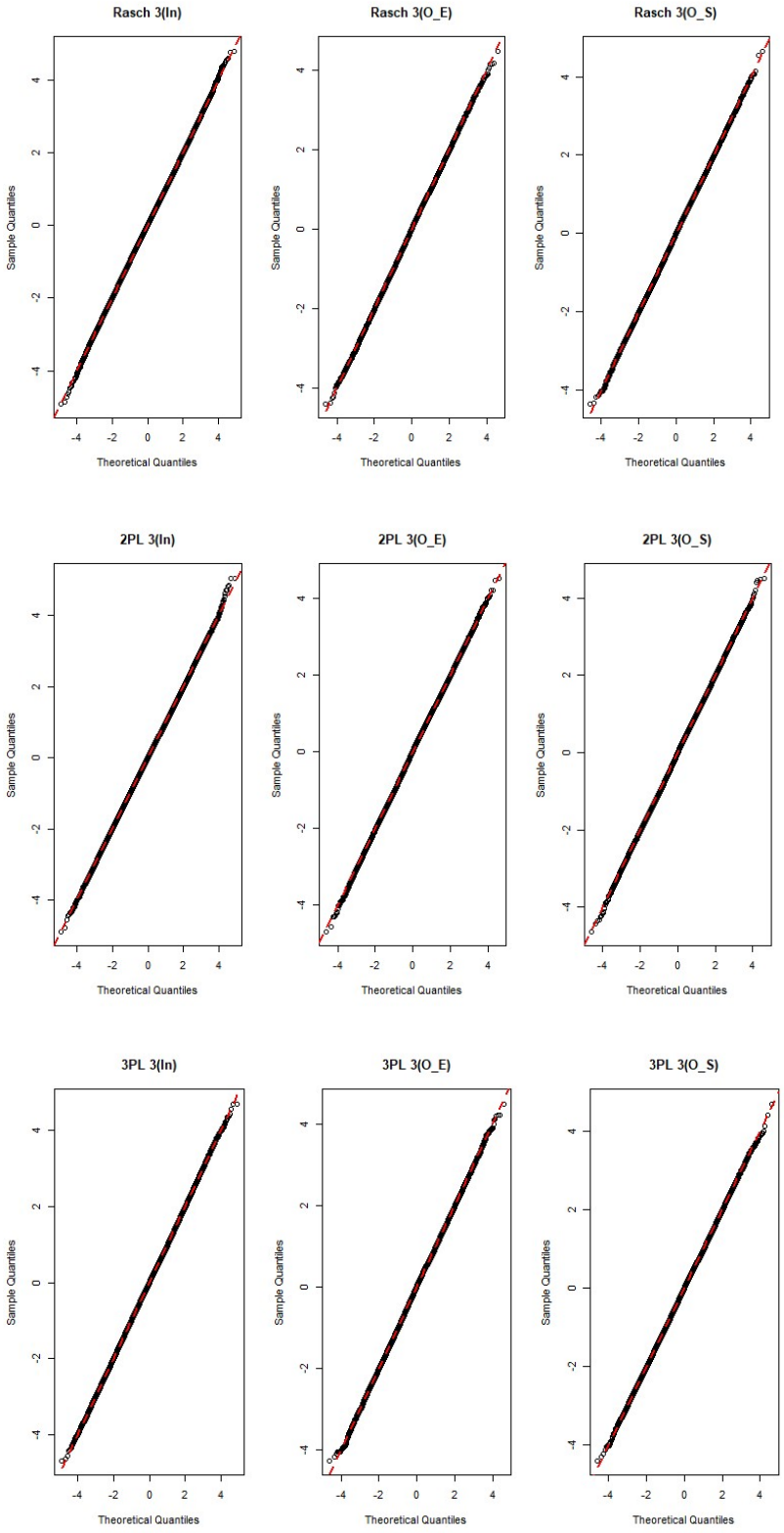


Figure C.3: RQR Checking Plot for Item 3.

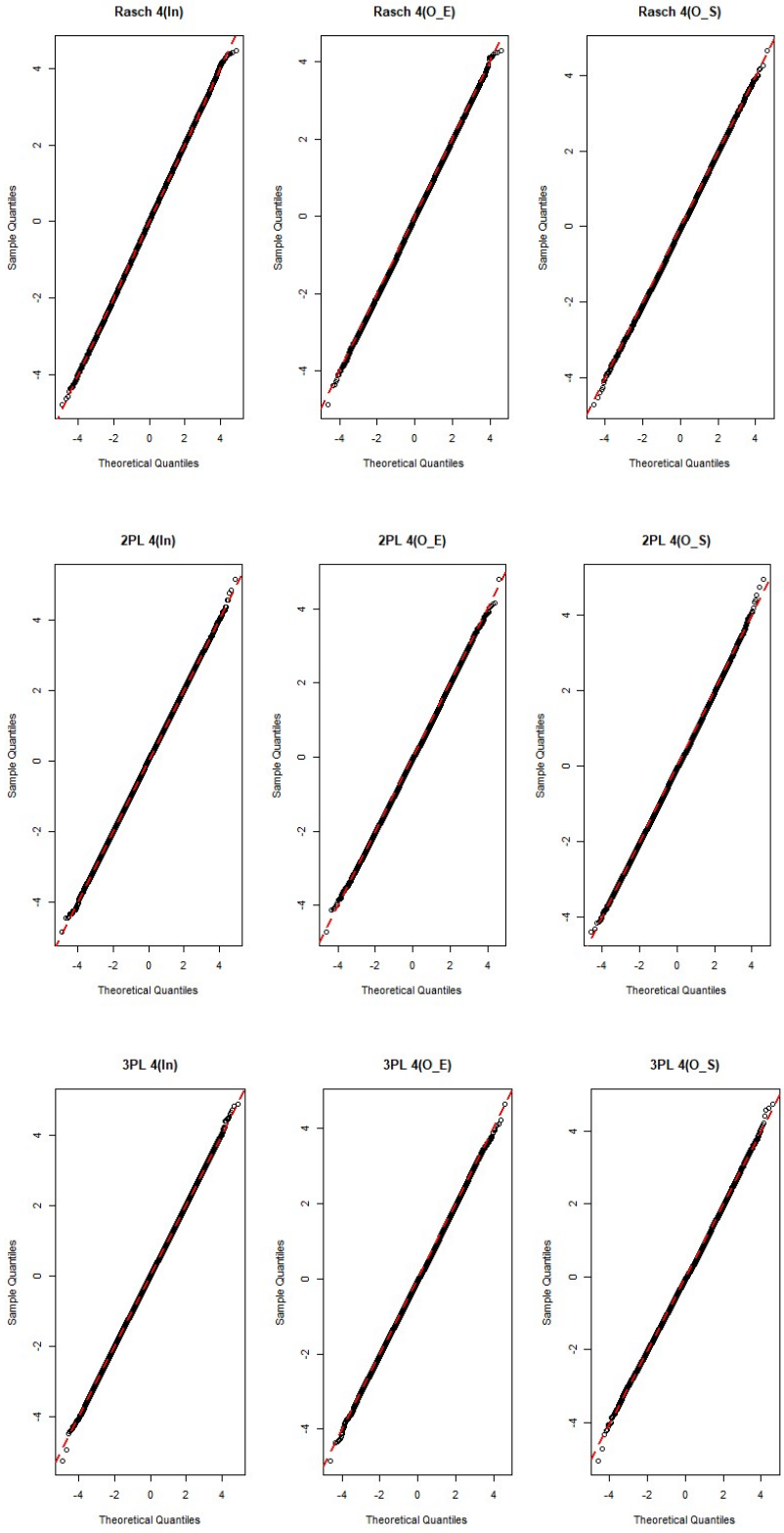


Figure C.4: RQR Checking Plot for Item 4.

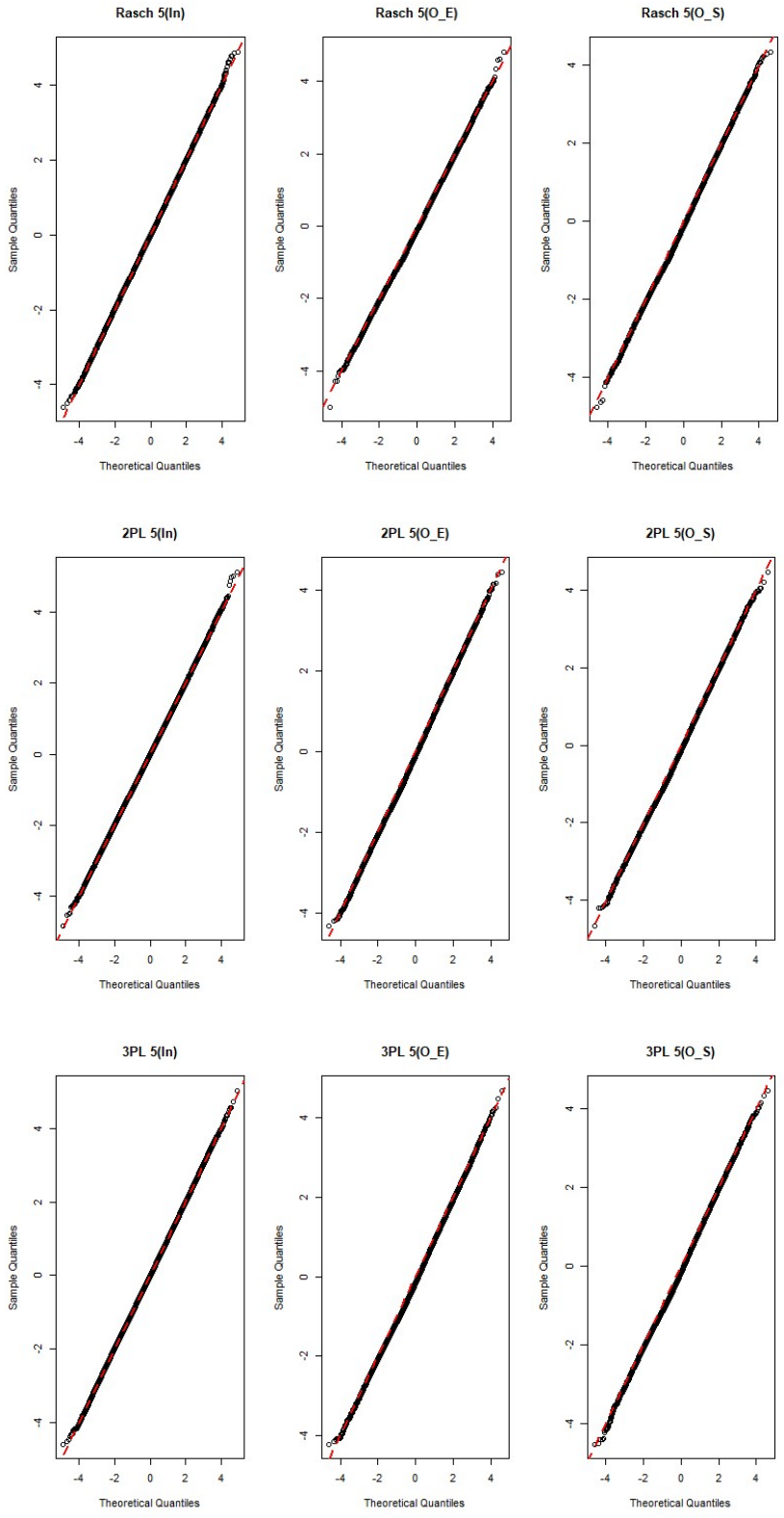


Figure C.5: RQR Checking Plot for Item 5.

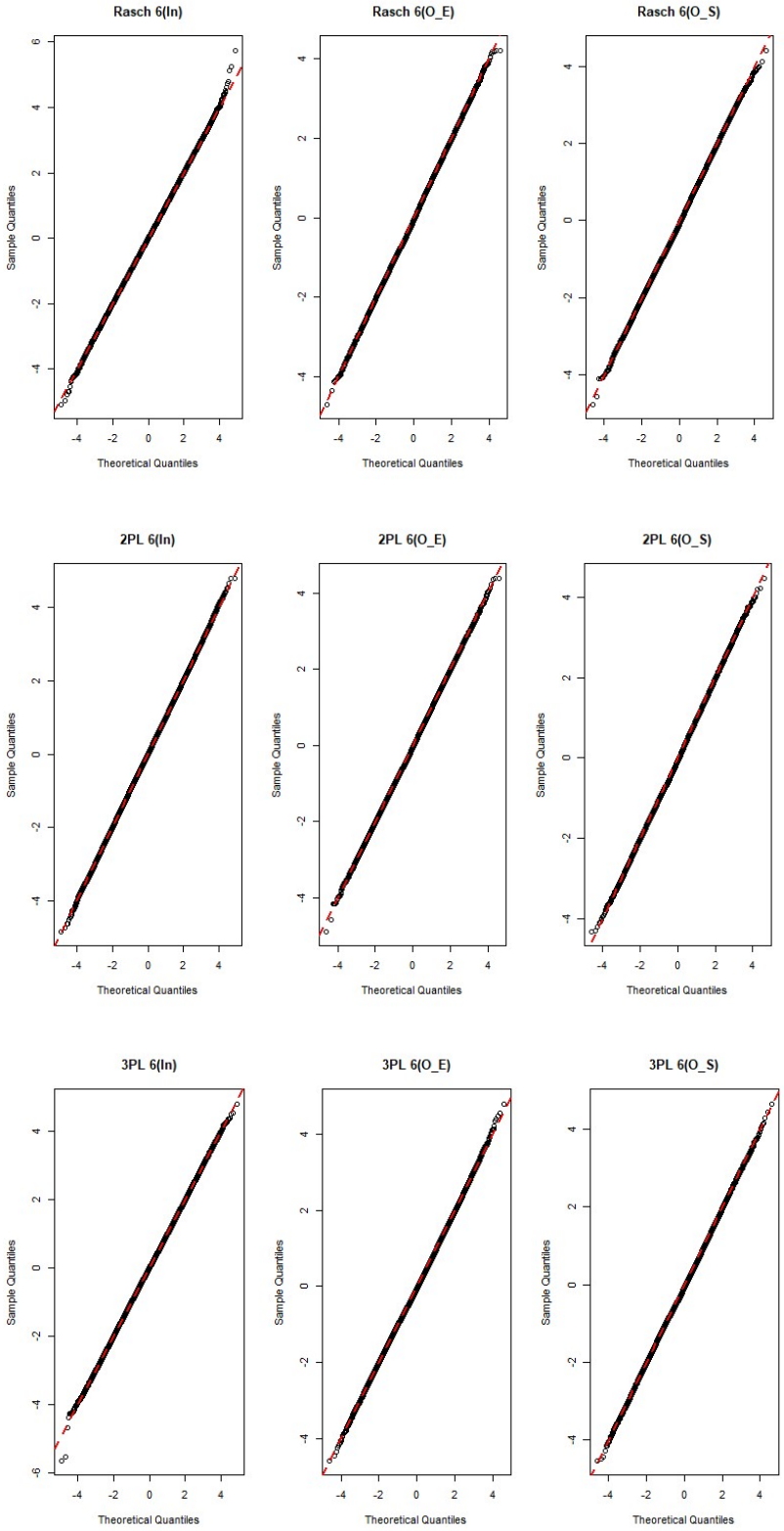


Figure C.6: RQR Checking Plot for Item 6.

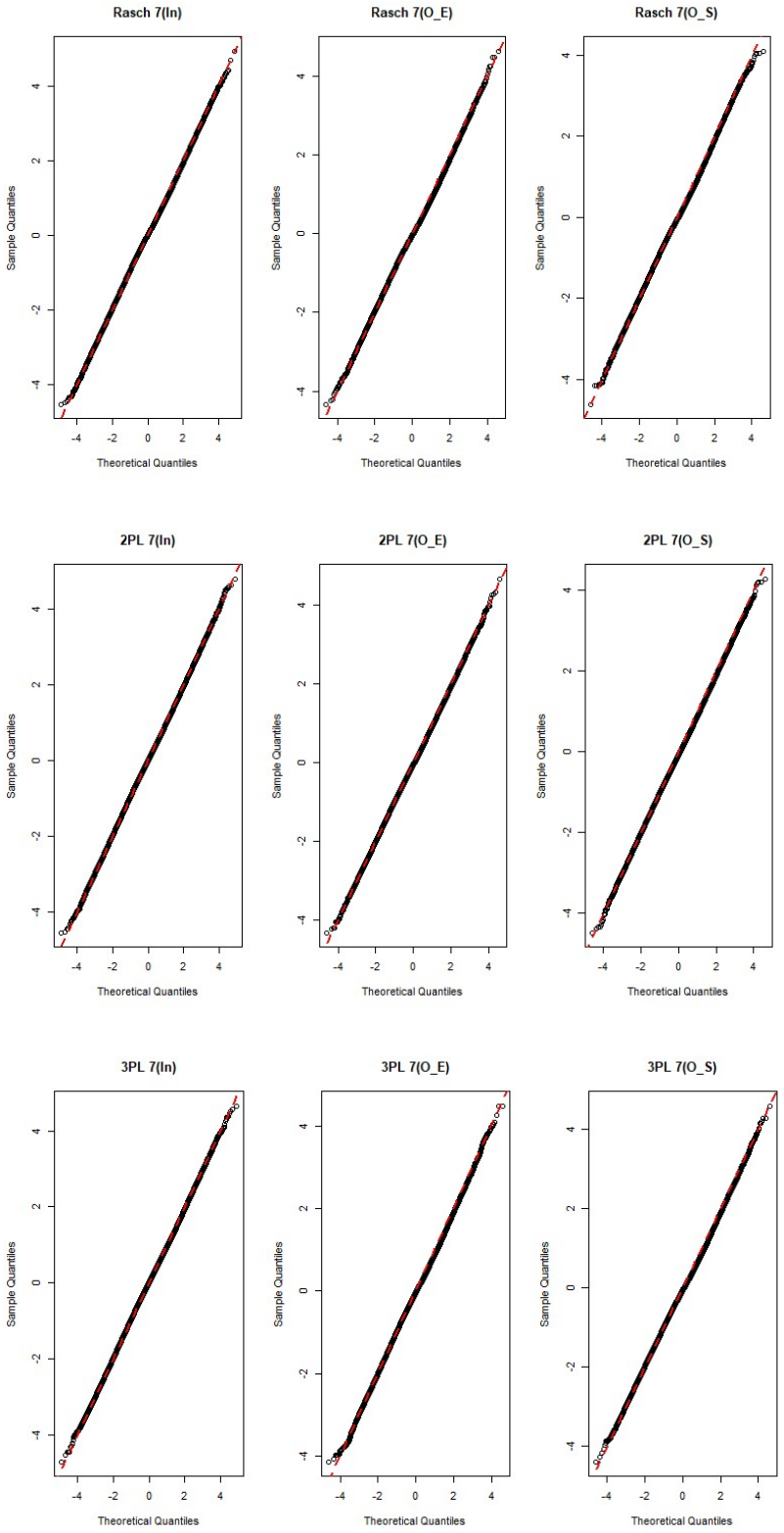


Figure C.7: RQR Checking Plot for Item 7.

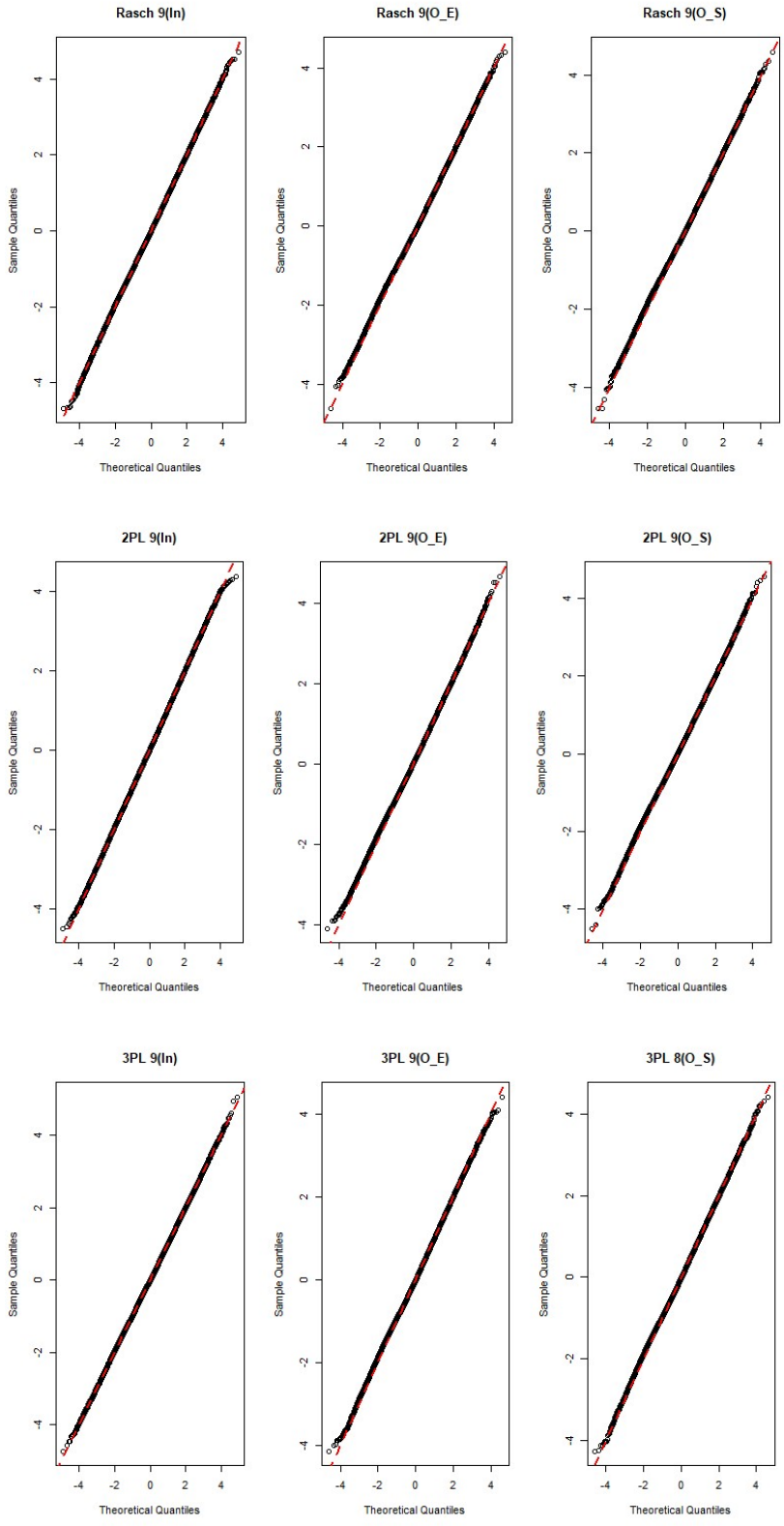


Figure C.8: RQR Checking Plot for Item 9.

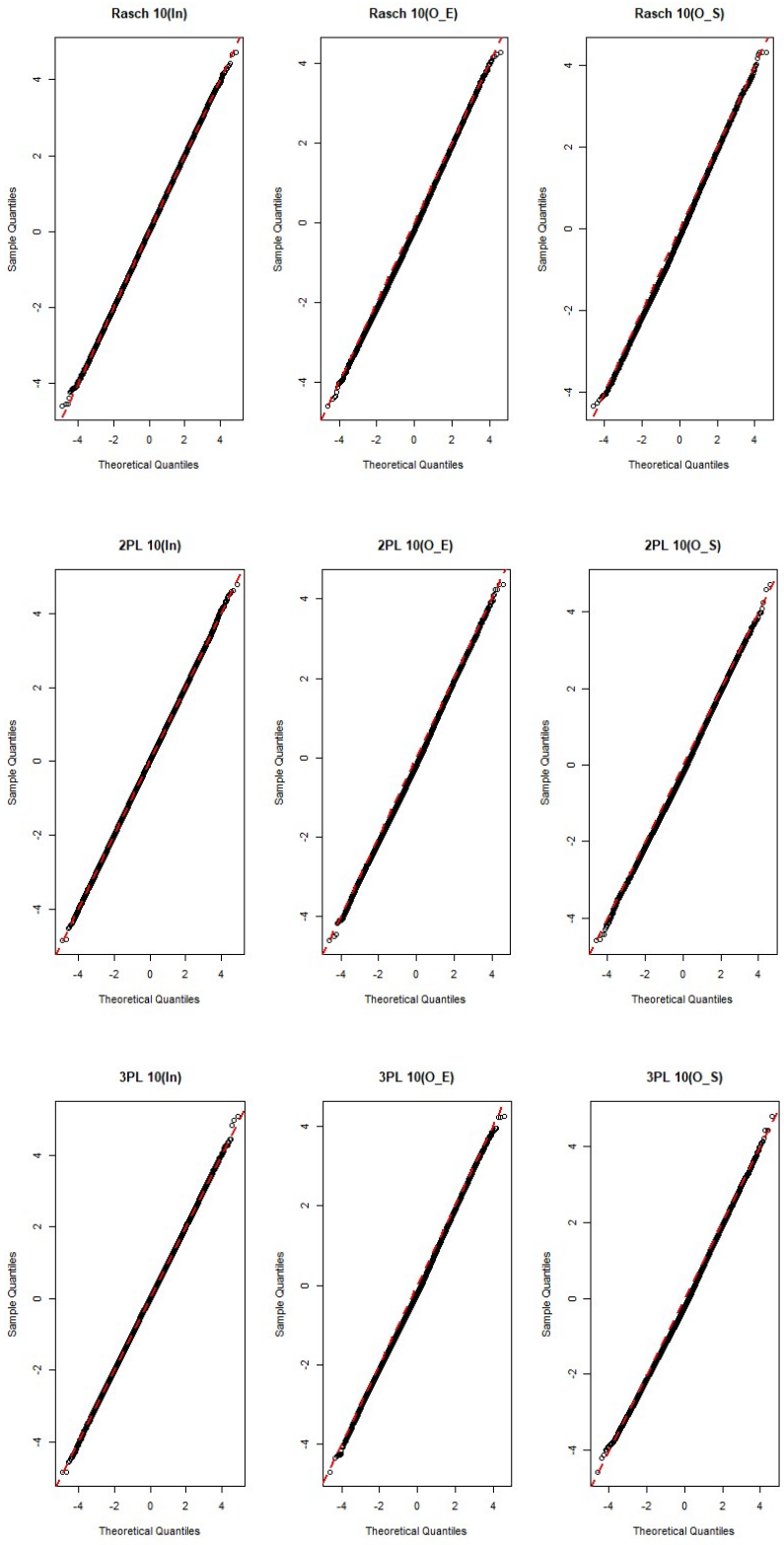


Figure C.9: RQR Checking Plot for Item 10.

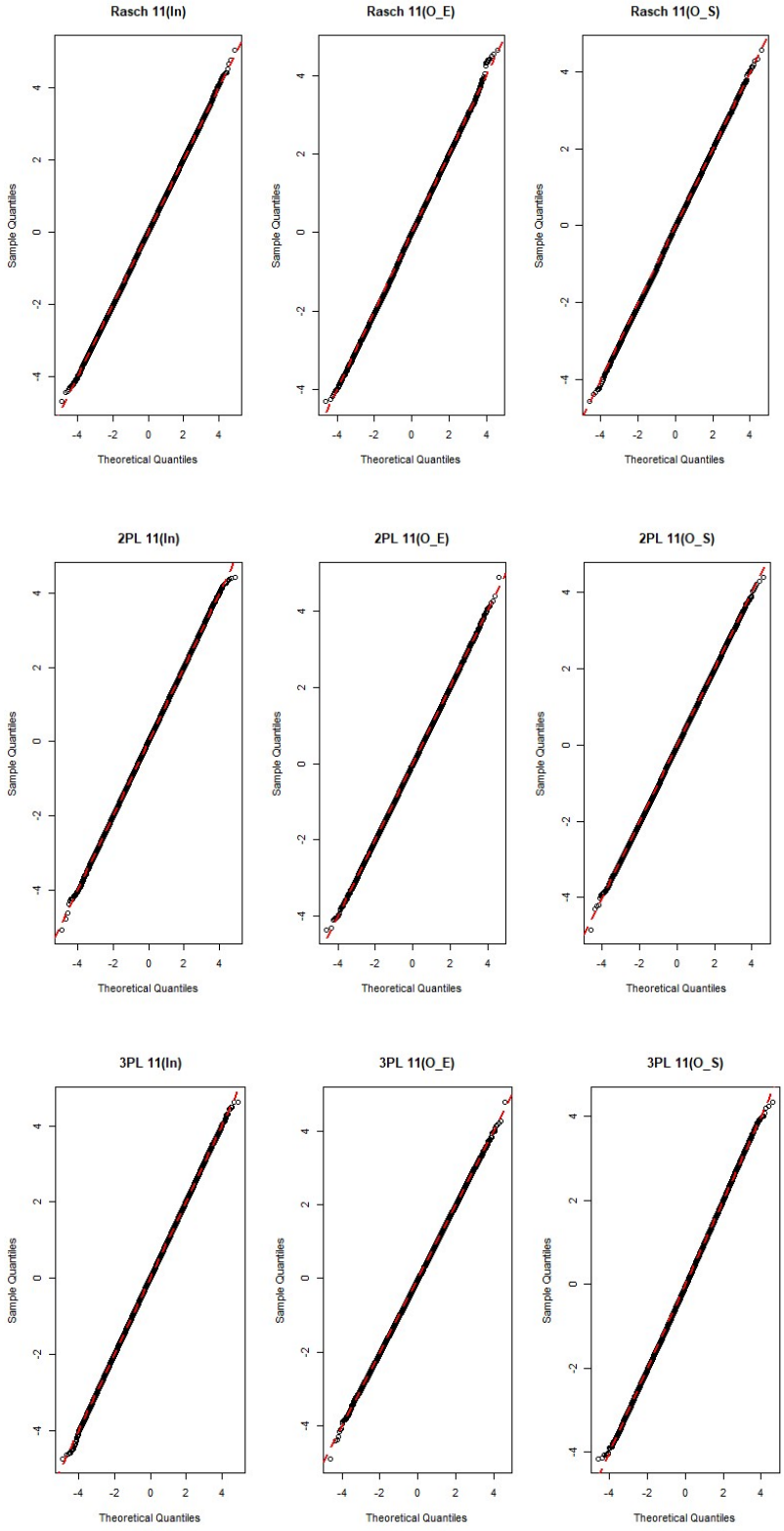


Figure C.10: RQR Checking Plot for Item 11.

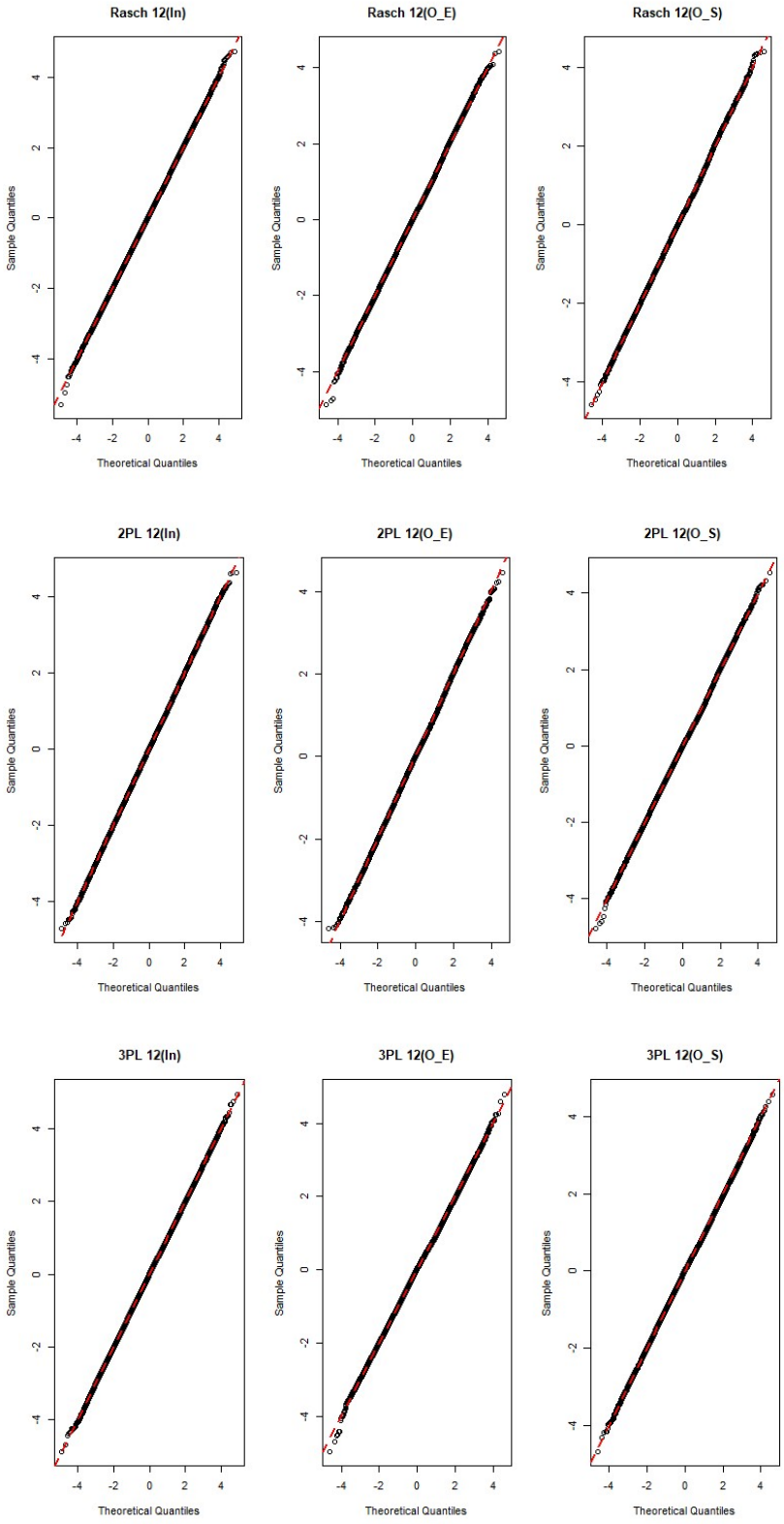


Figure C.11: RQR Checking Plot for Item 12.

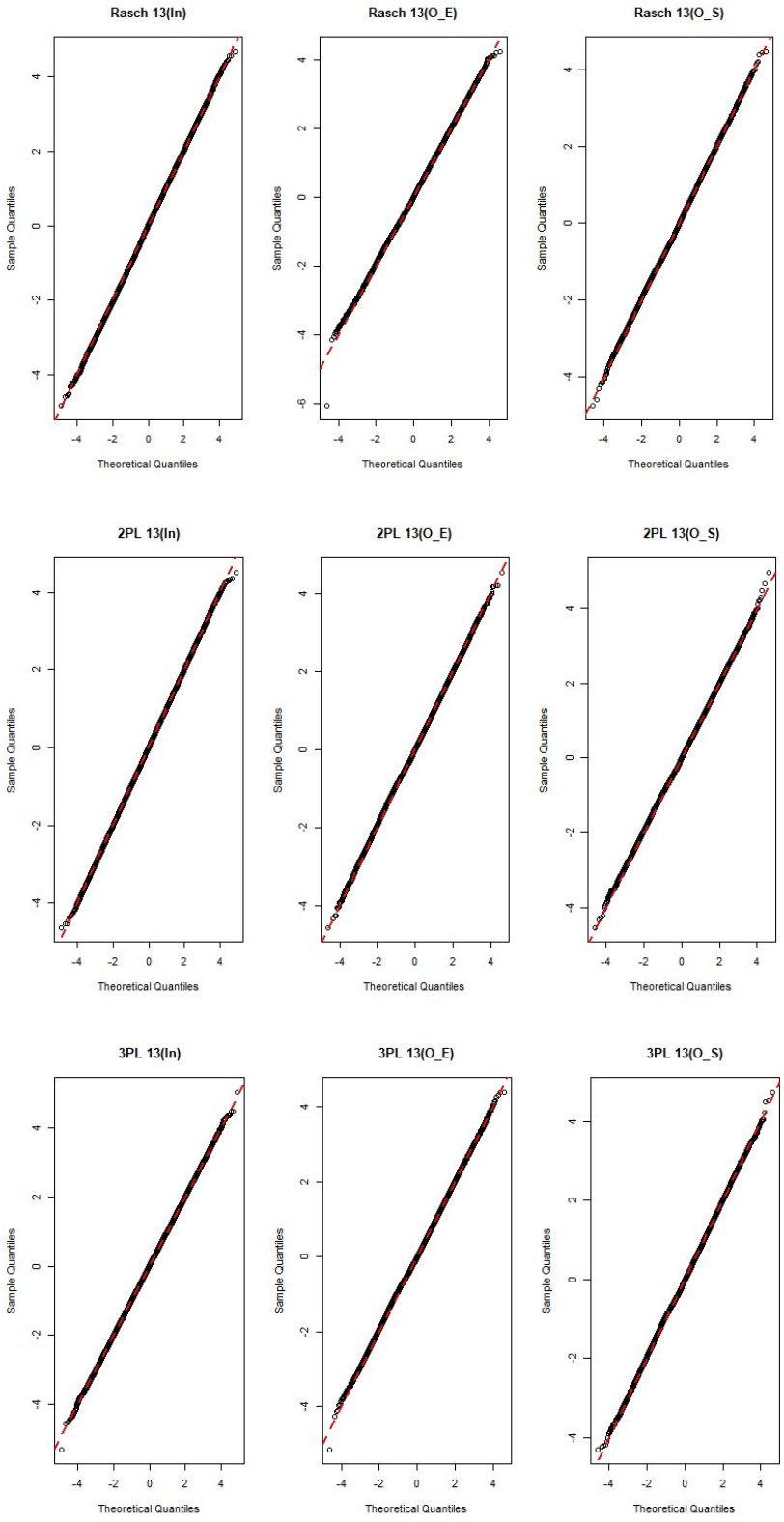


Figure C.12: RQR Checking Plot for Item 13.

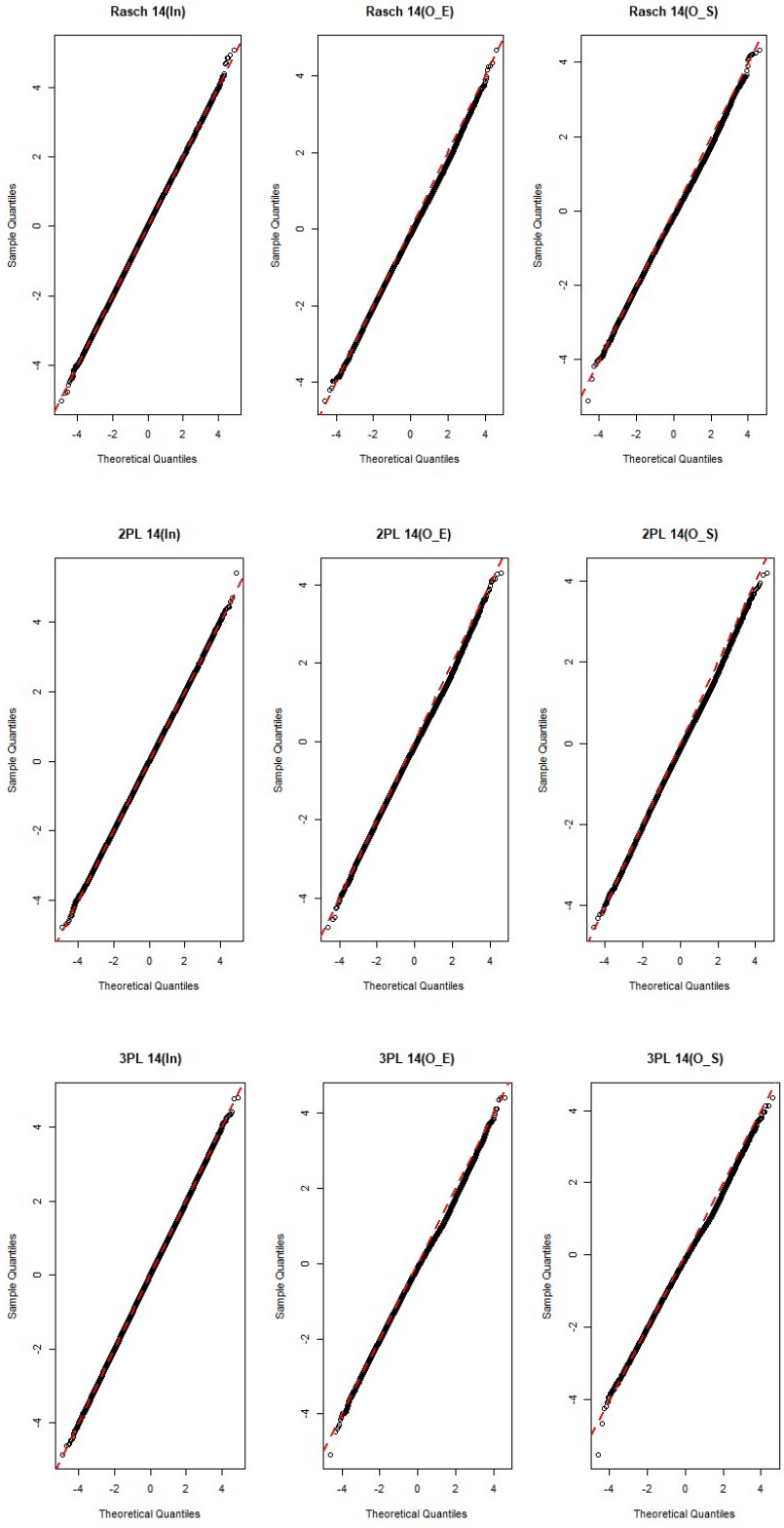


Figure C.13: RQR Checking Plot for Item 14.

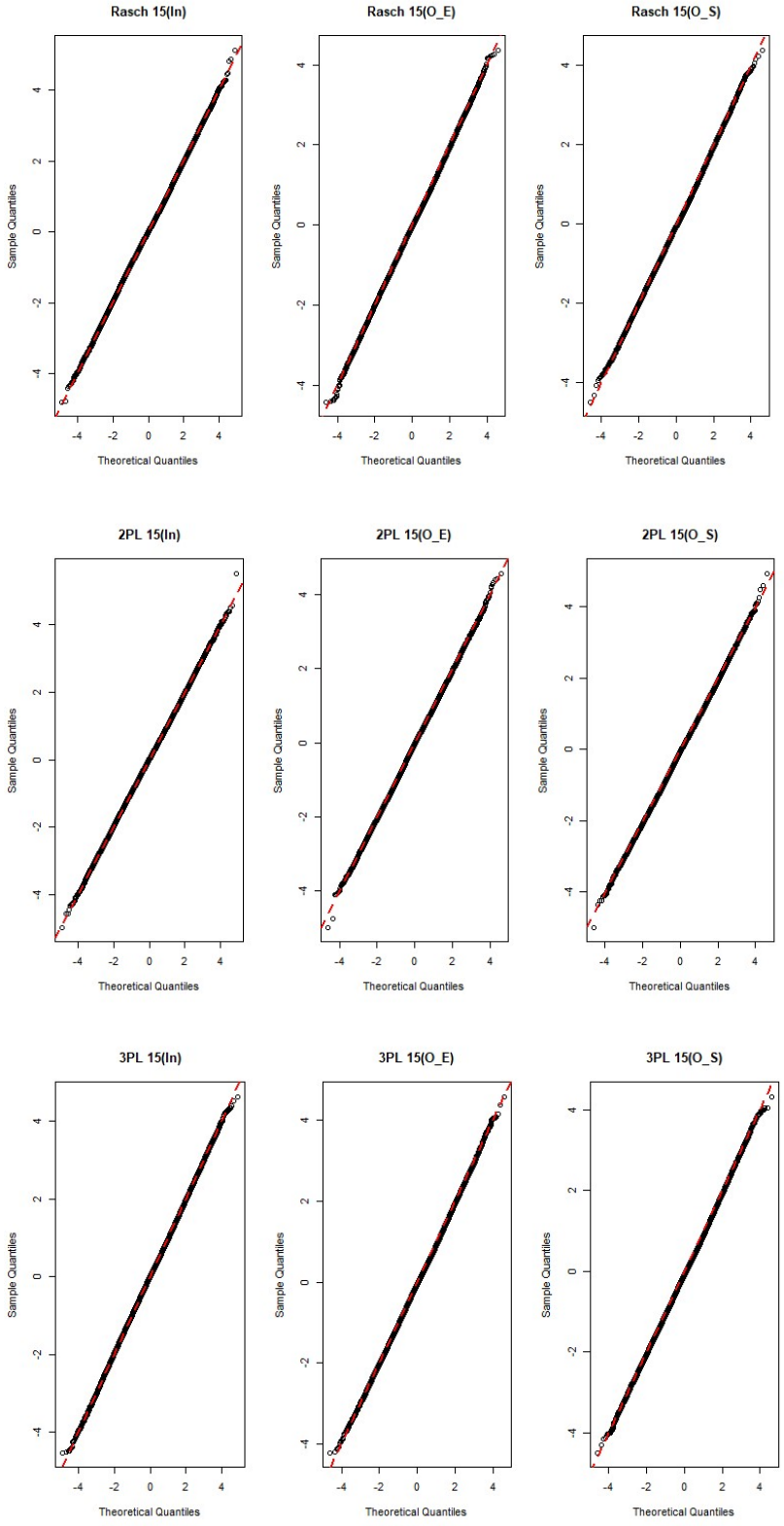


Figure C.14: RQR Checking Plot for Item 15.

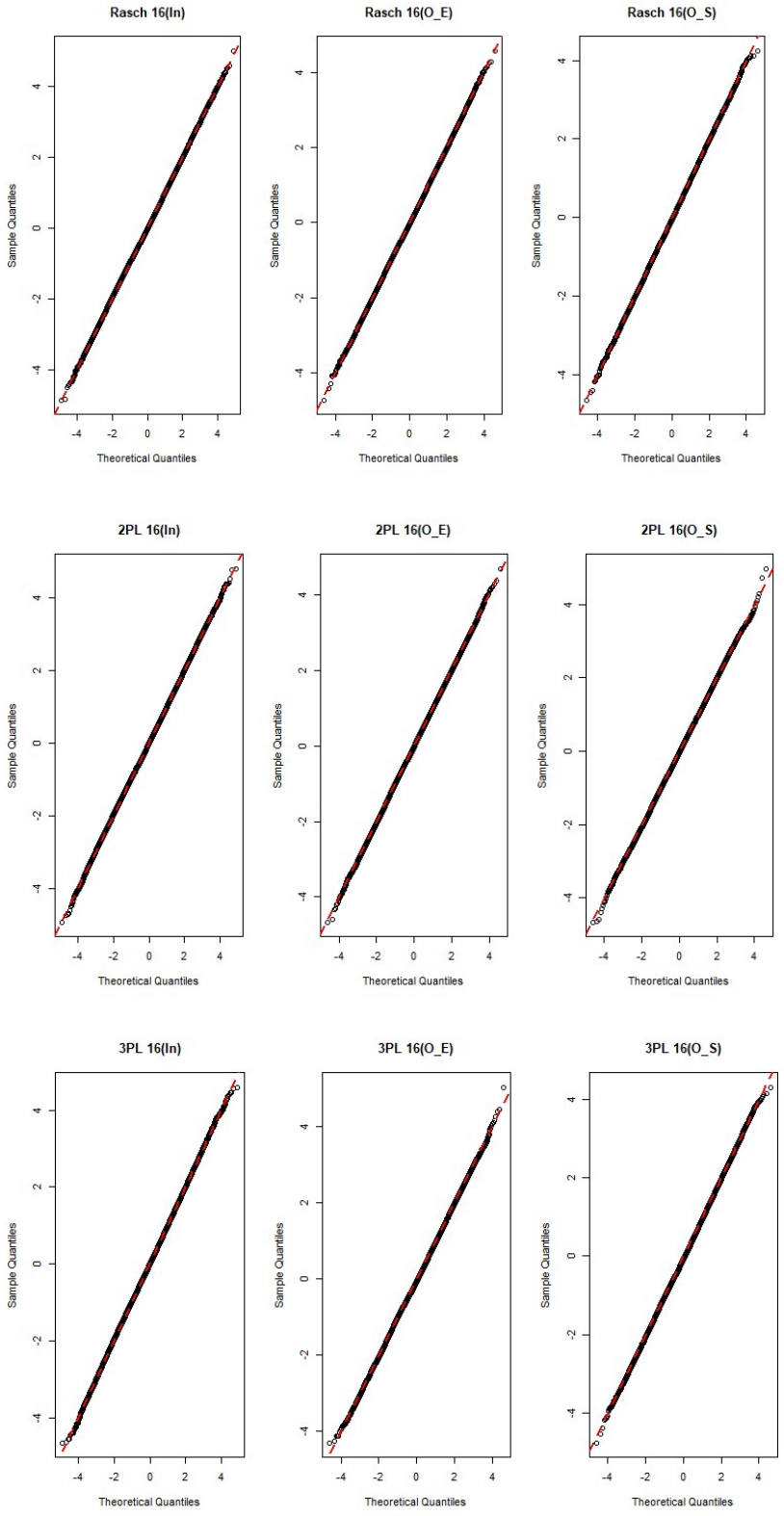


Figure C.15: RQR Checking Plot for Item 16.

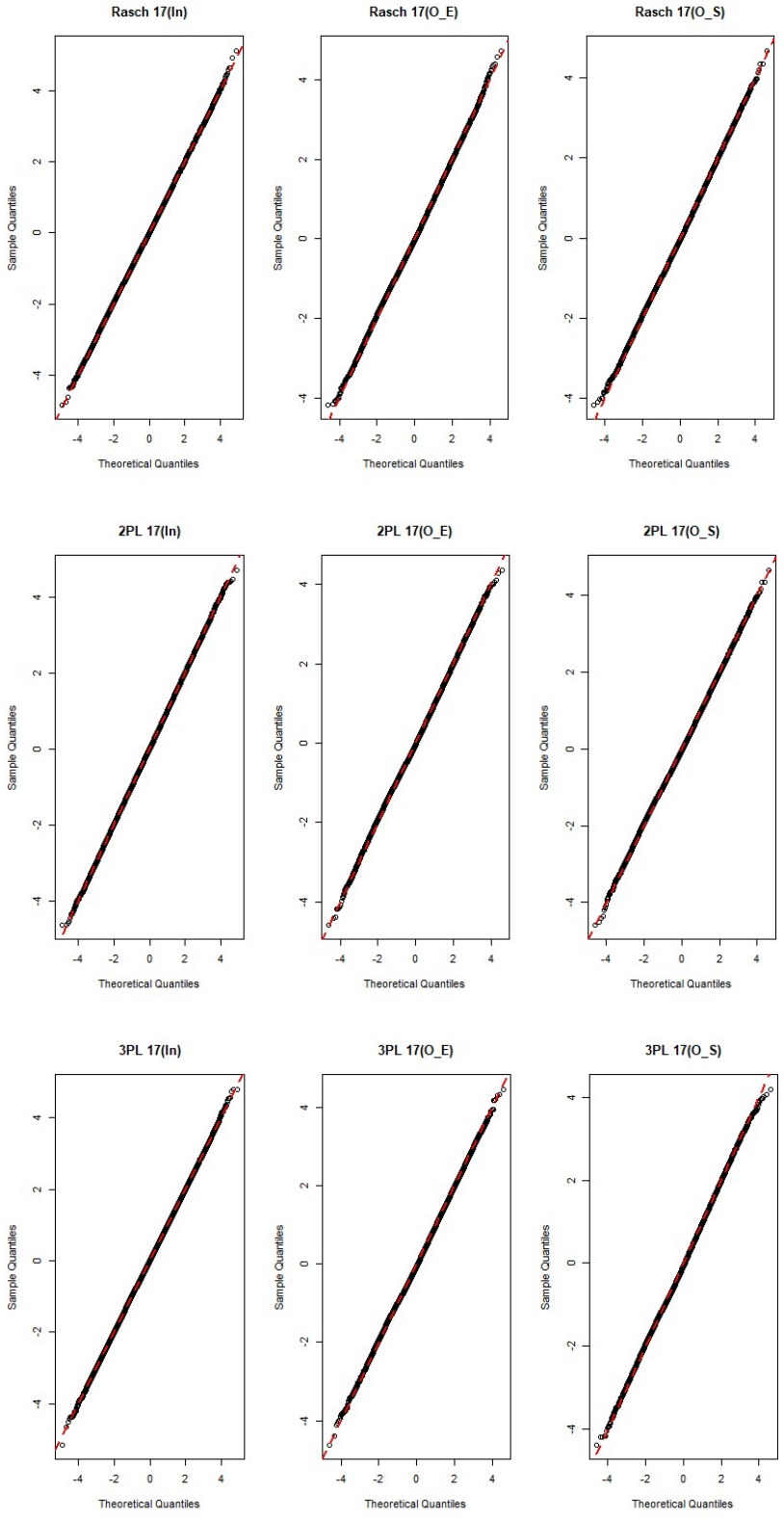


Figure C.16: RQR Checking Plot for Item 17.

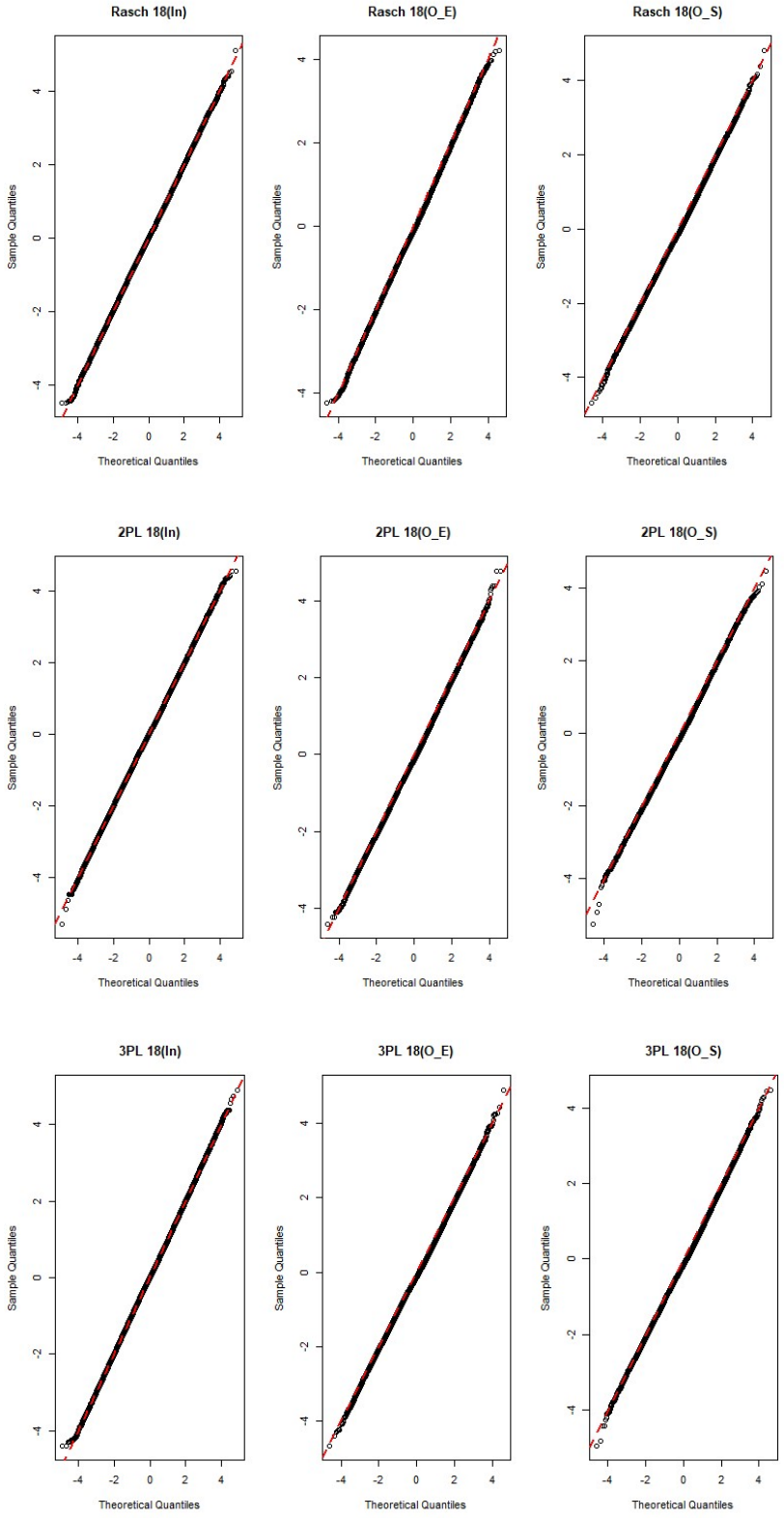


Figure C.17: RQR Checking Plot for Item 18.

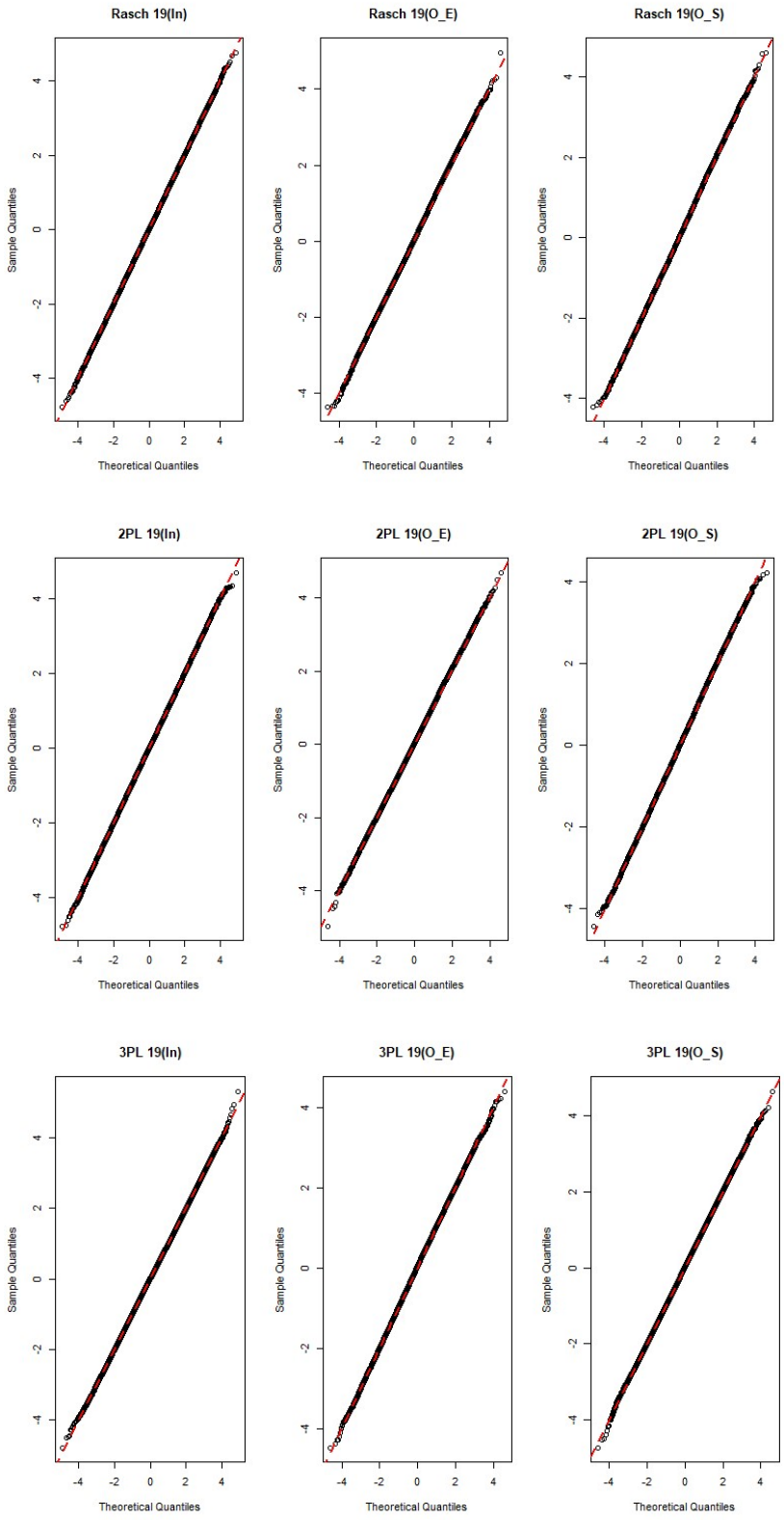


Figure C.18: RQR Checking Plot for Item 19.

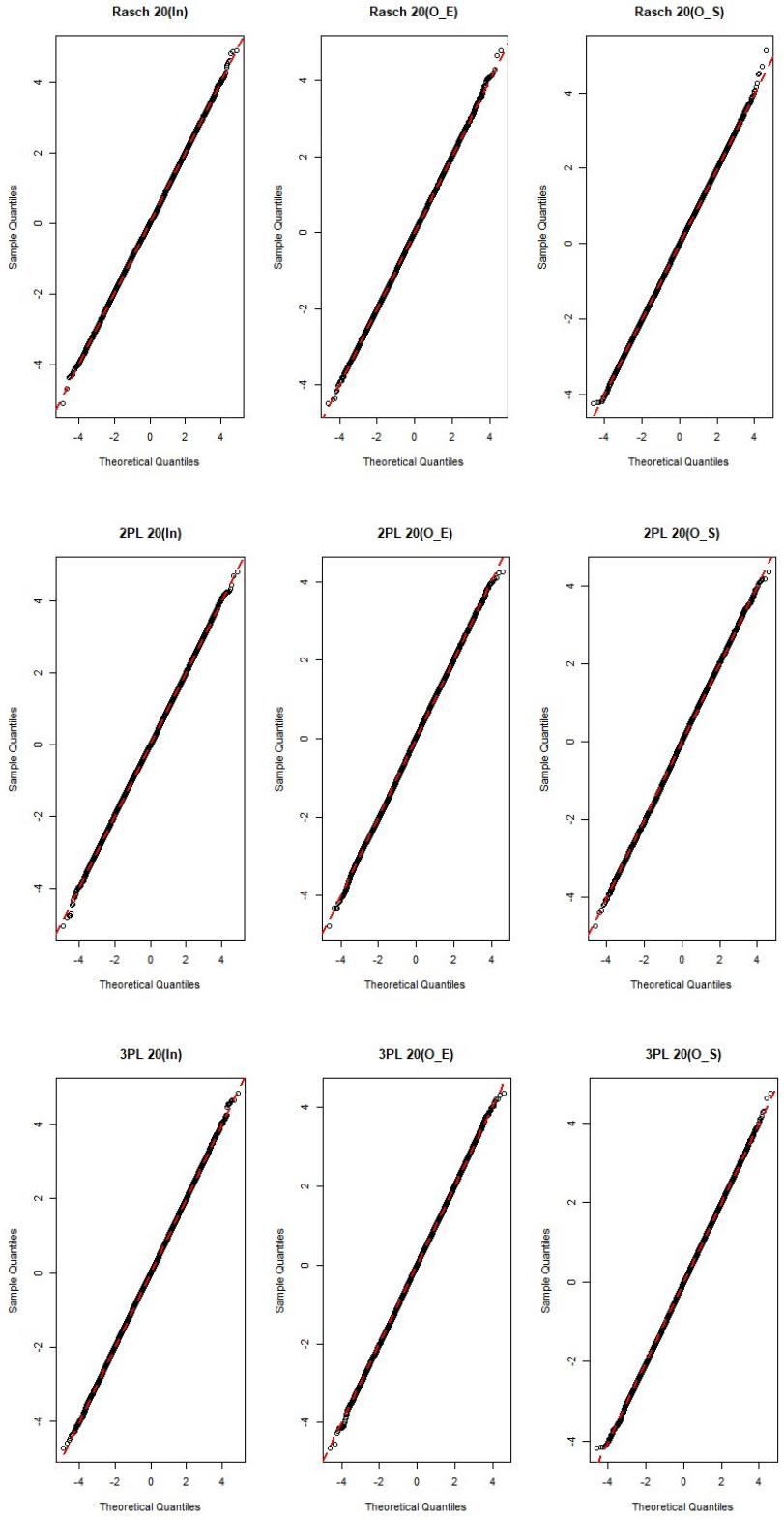


Figure C.19: RQR Checking Plot for Item 20.

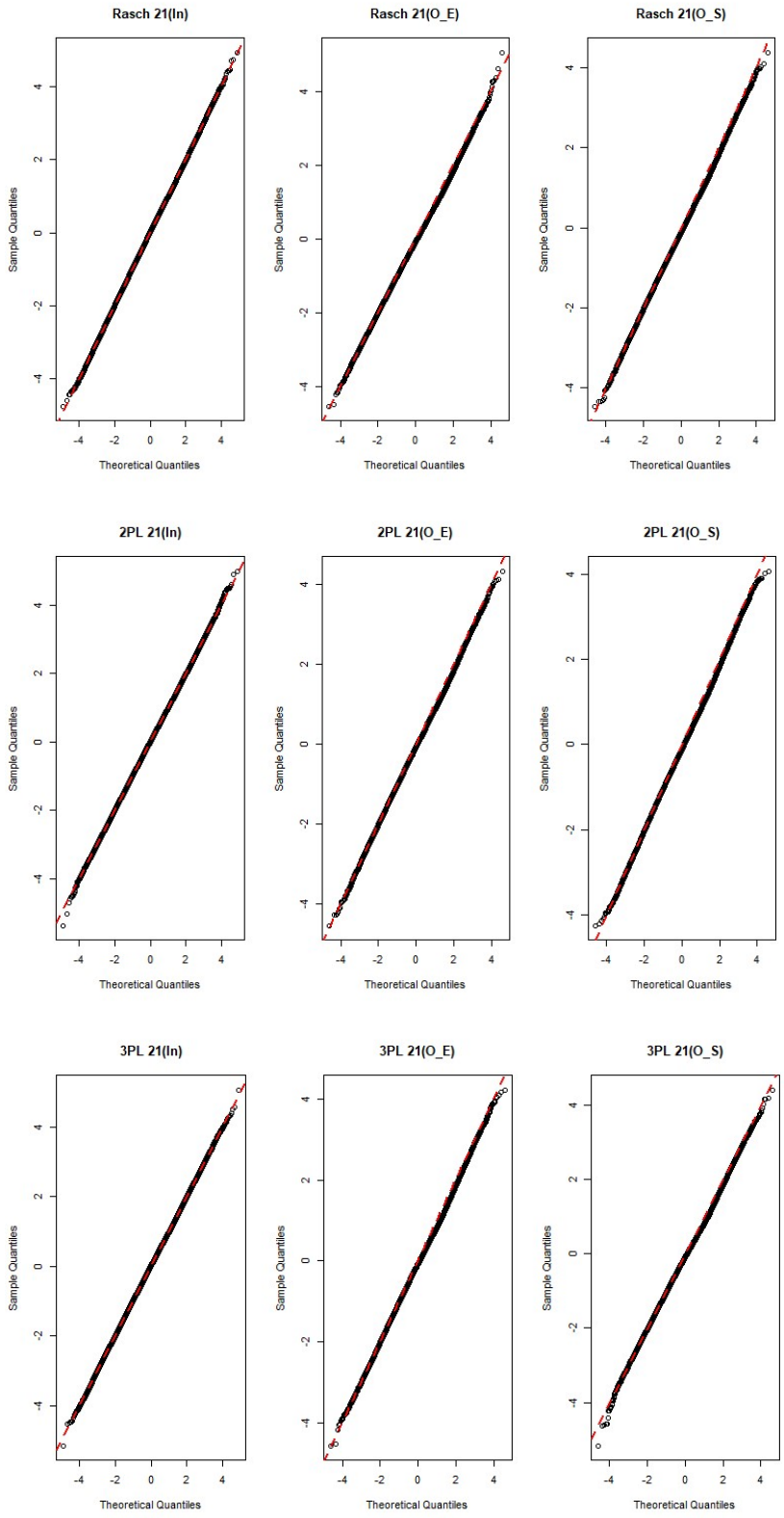


Figure C.20: RQR Checking Plot for Item 21.

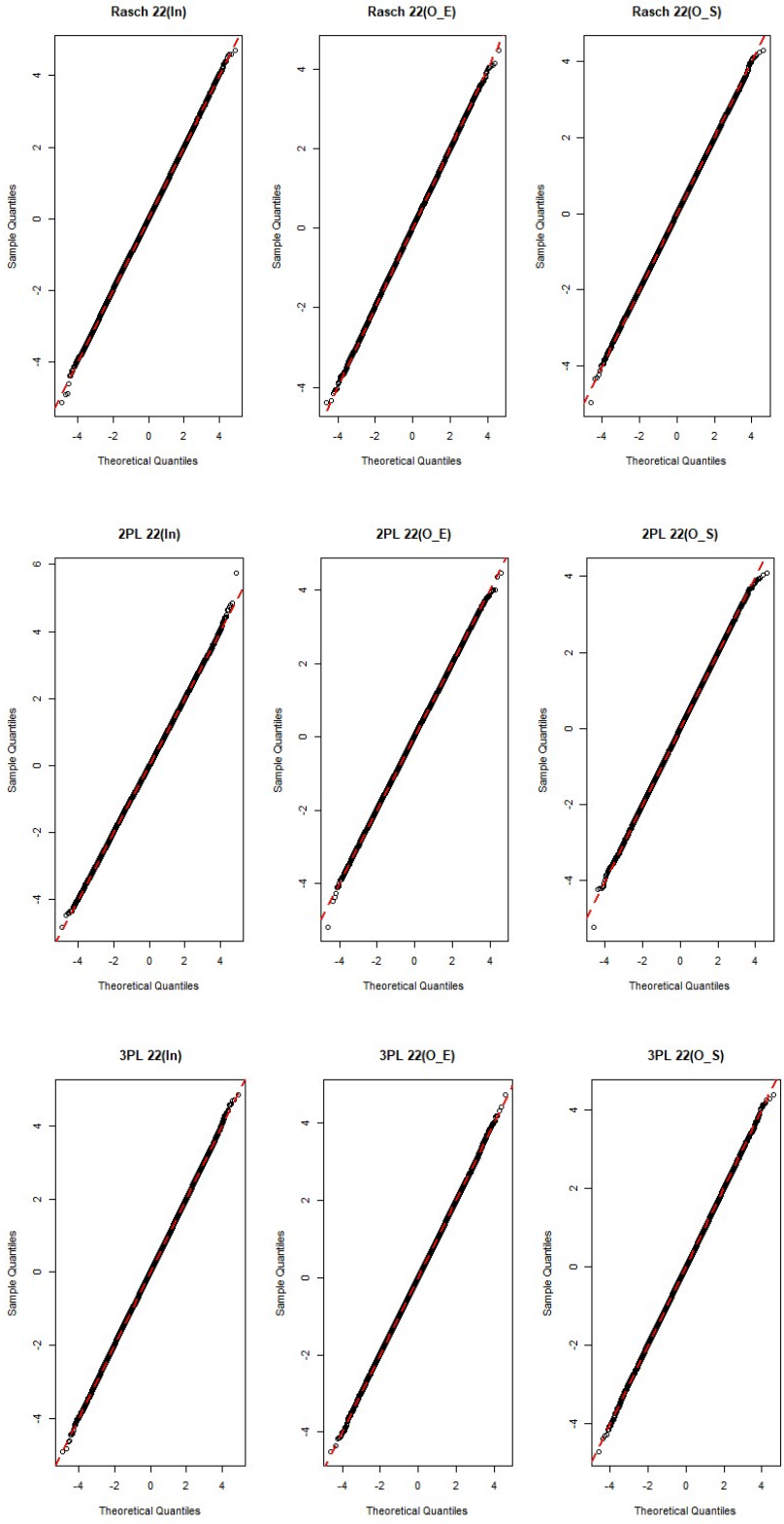


Figure C.21: RQR Checking Plot for Item 22.

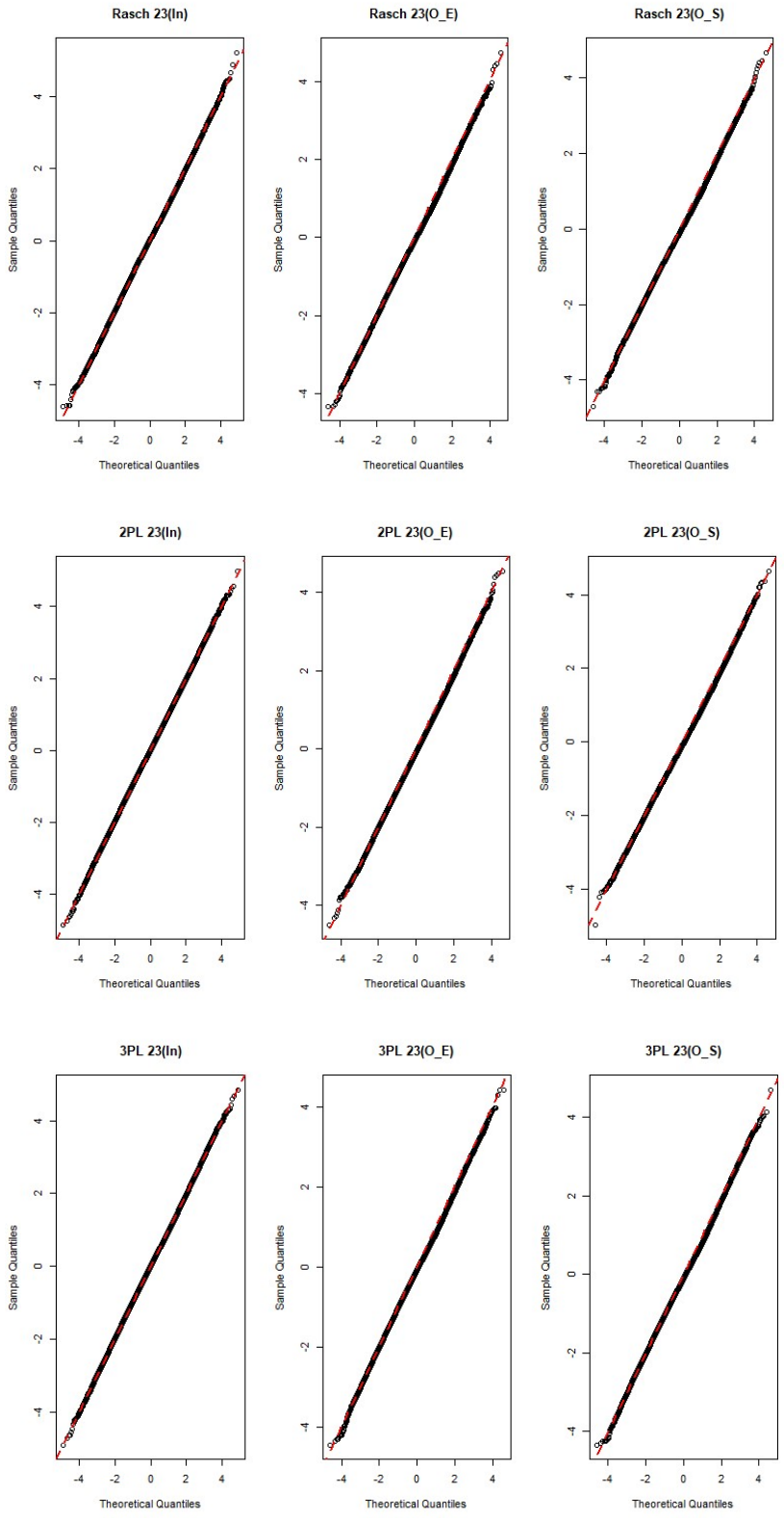


Figure C.22: RQR Checking Plot for Item 23.

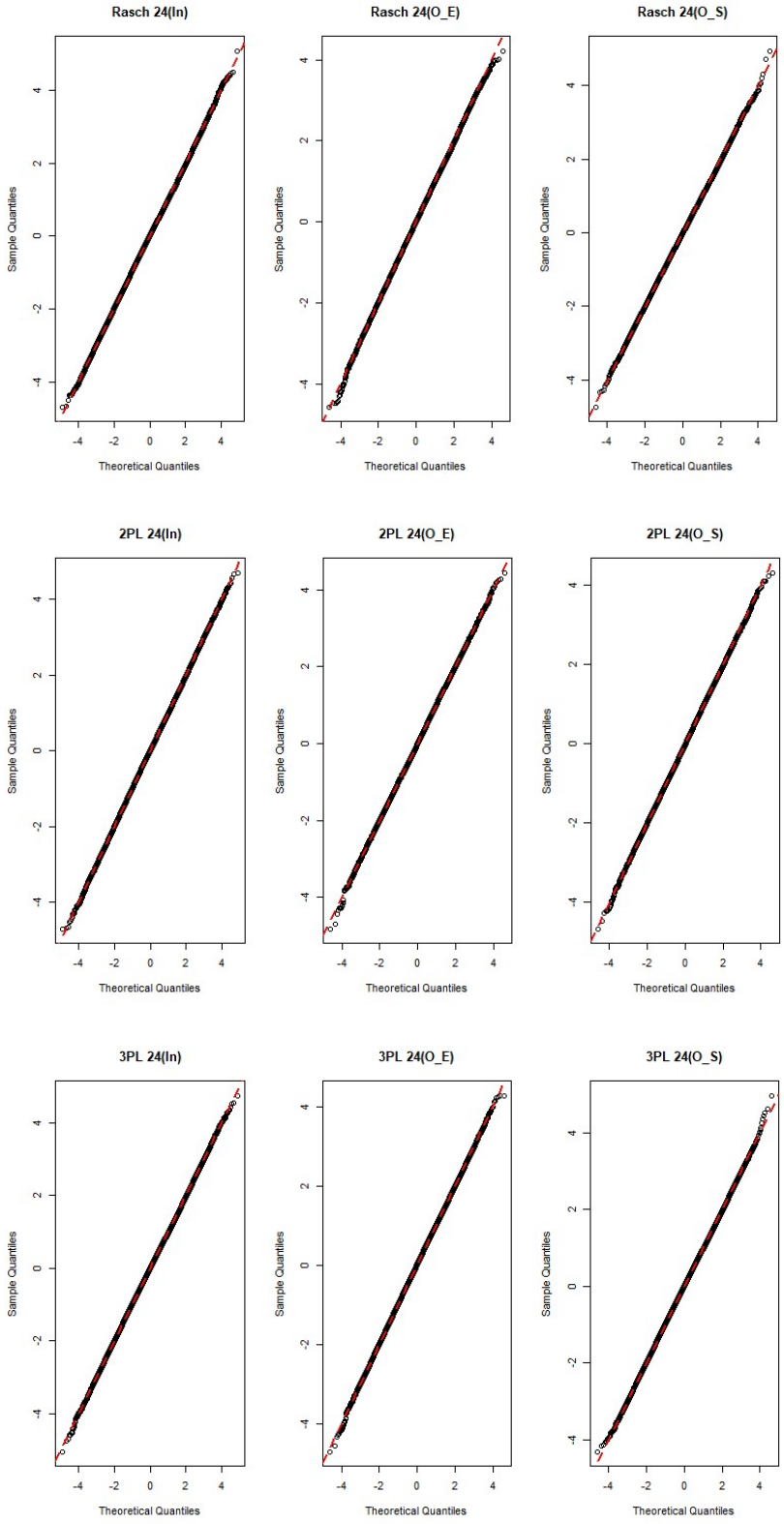


Figure C.23: RQR Checking Plot for Item 24.

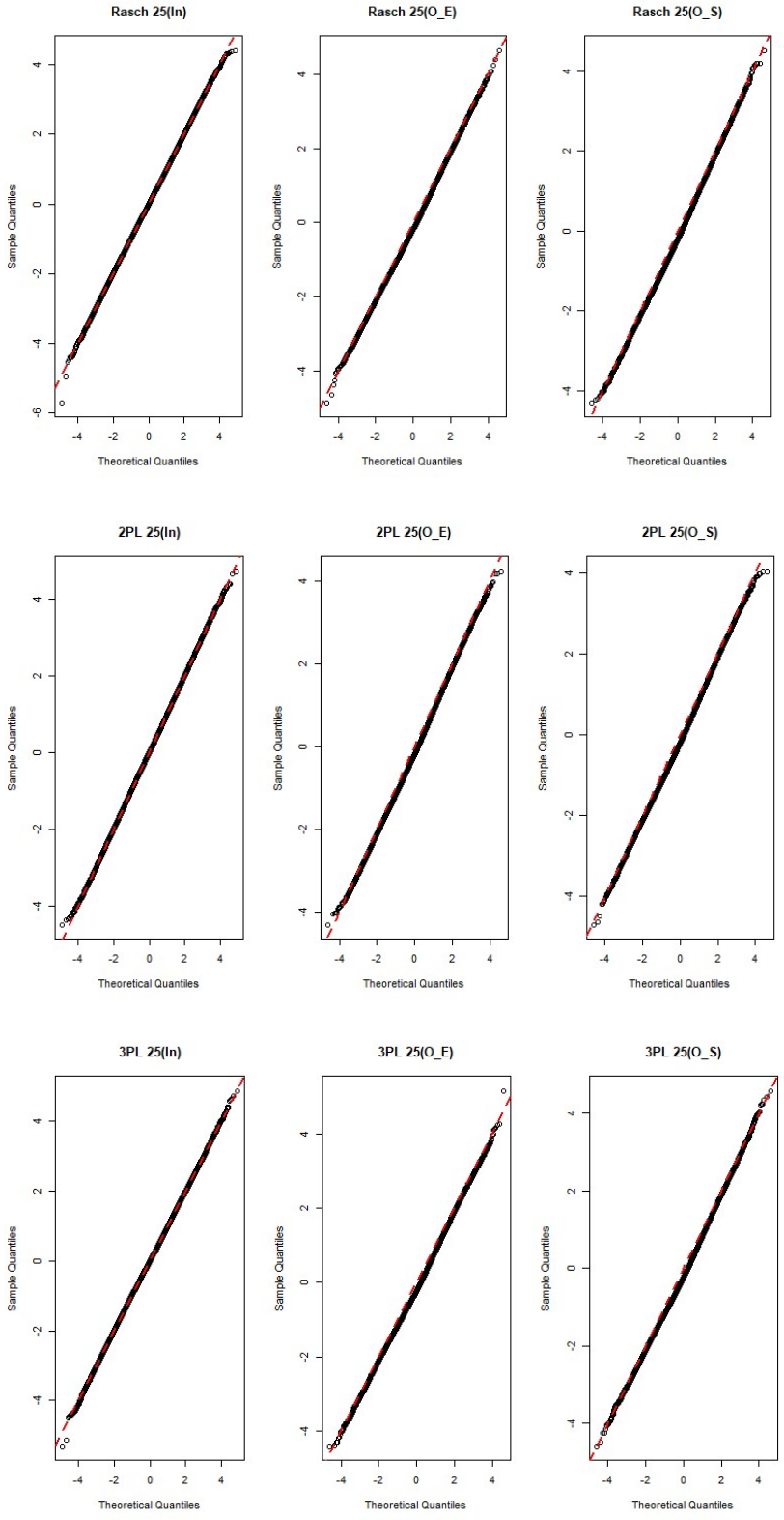


Figure C.24: RQR Checking Plot for Item 25.

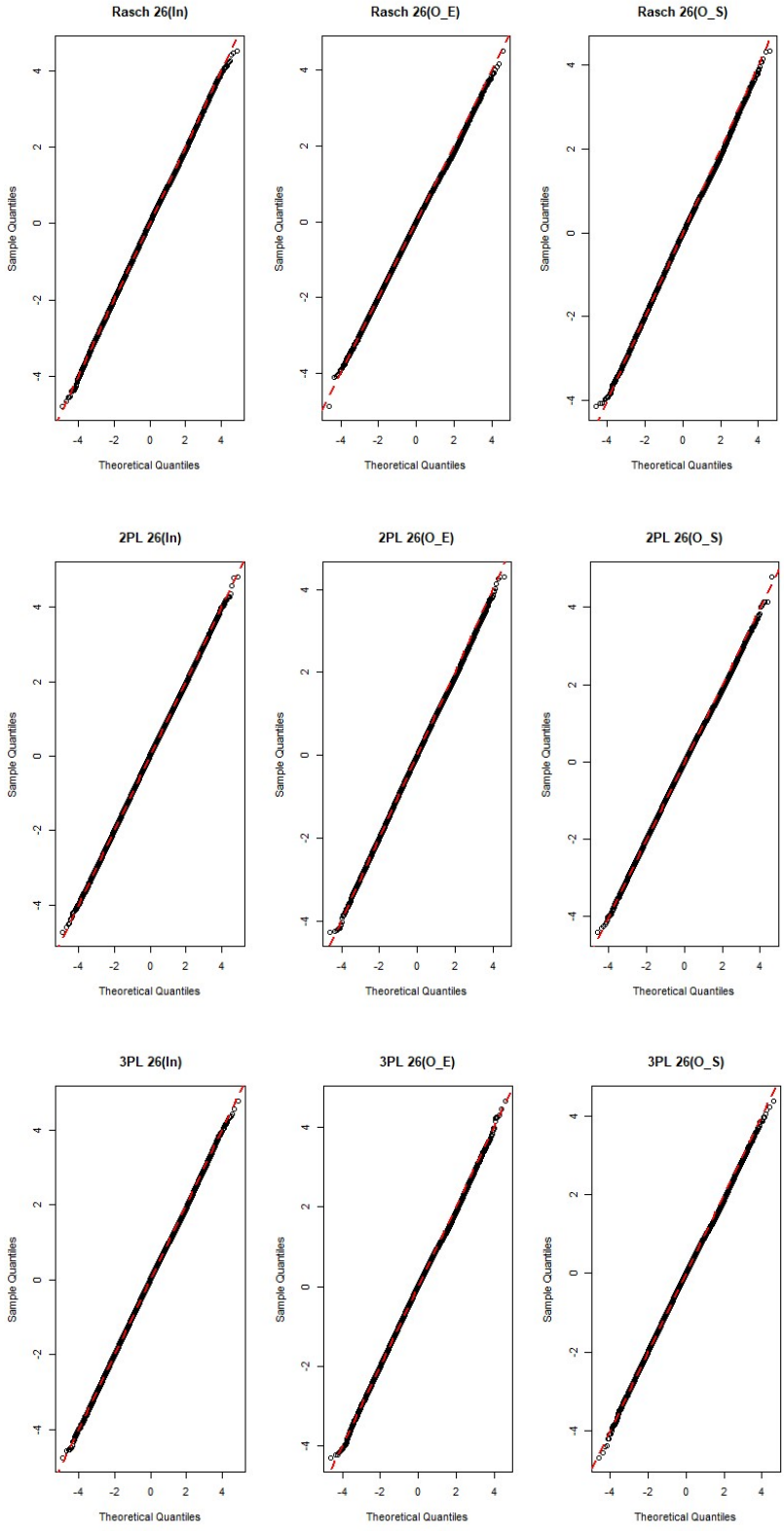


Figure C.25: RQR Checking Plot for Item 26.

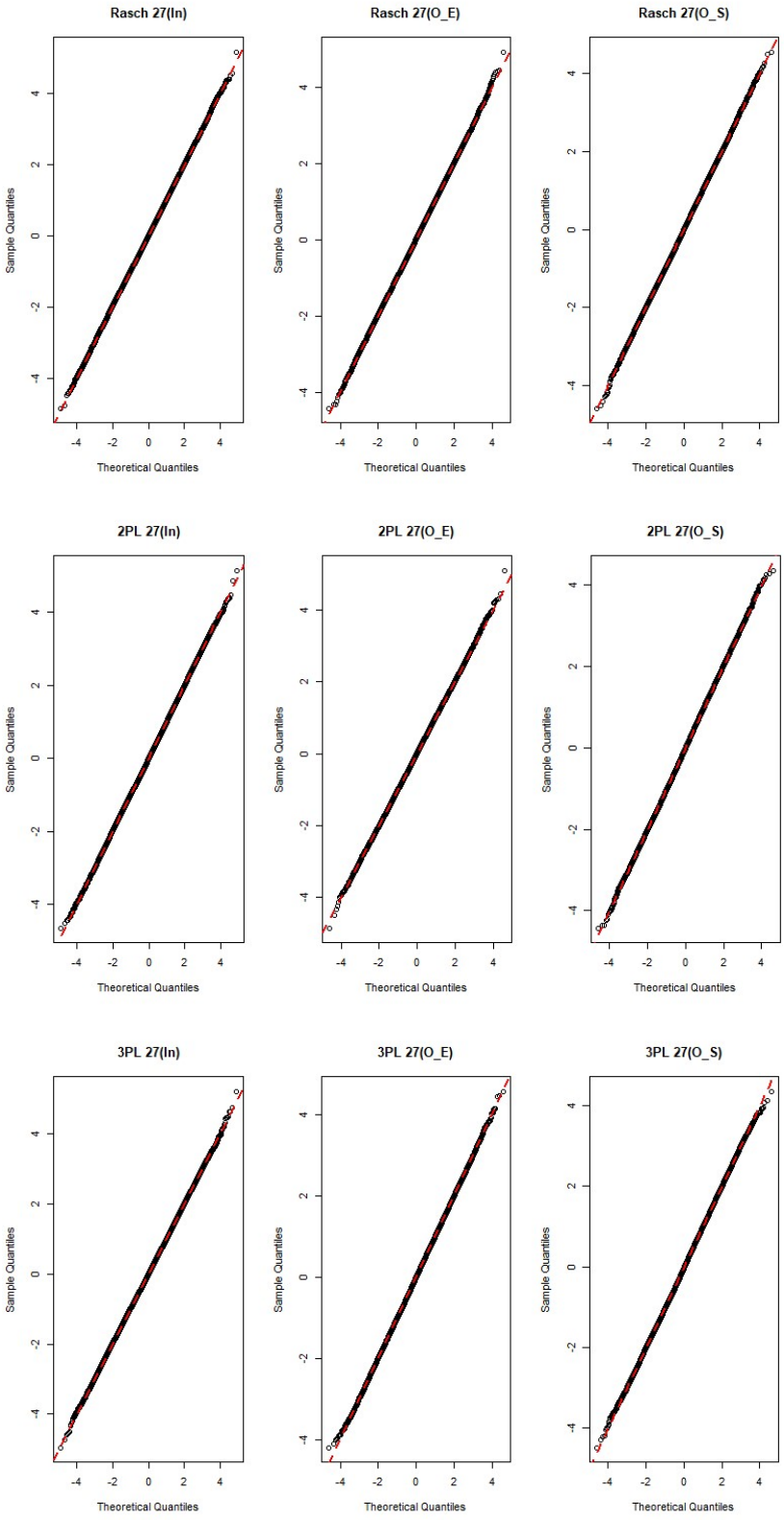


Figure C.26: RQR Checking Plot for Item 27.

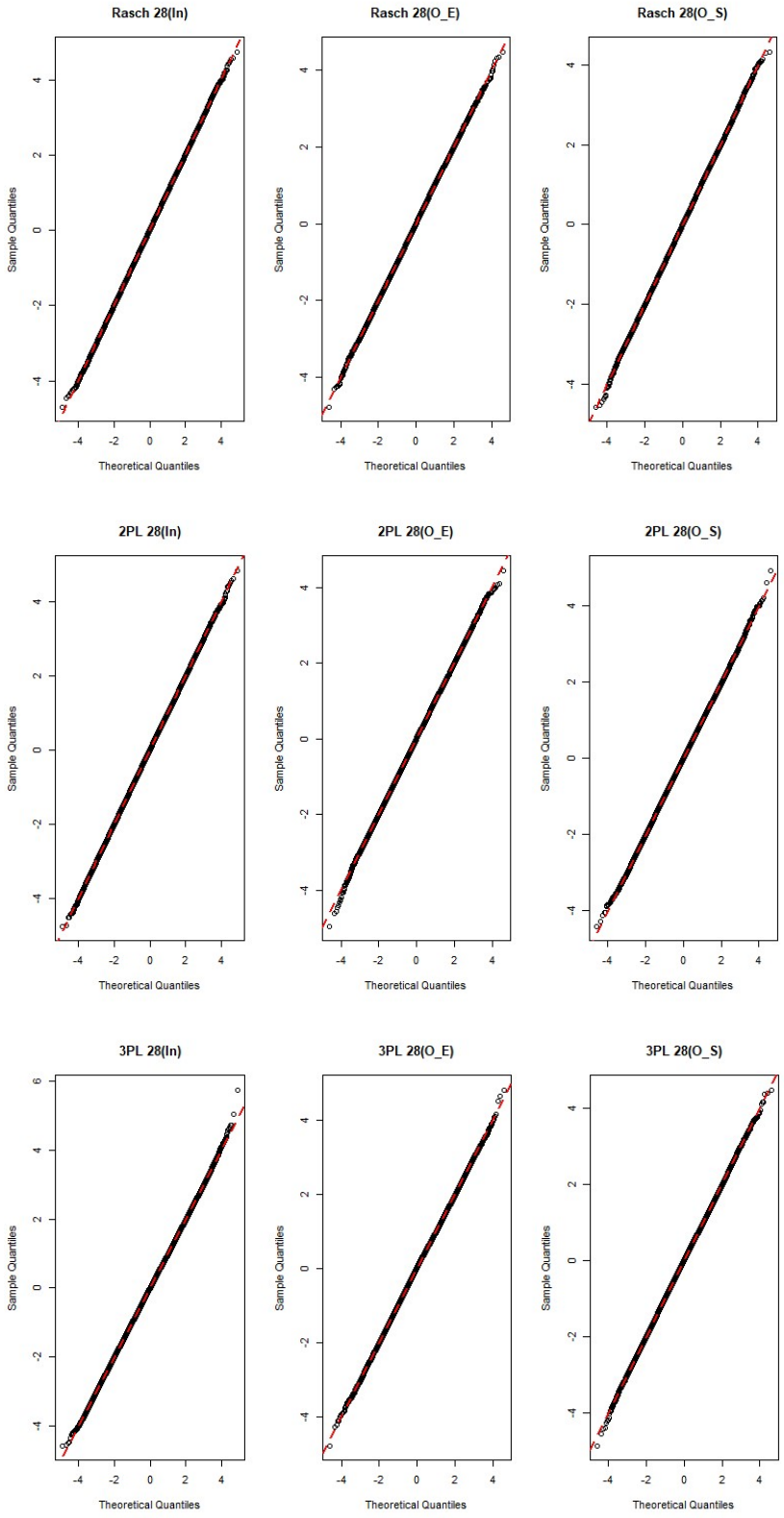


Figure C.27: RQR Checking Plot for Item 28.

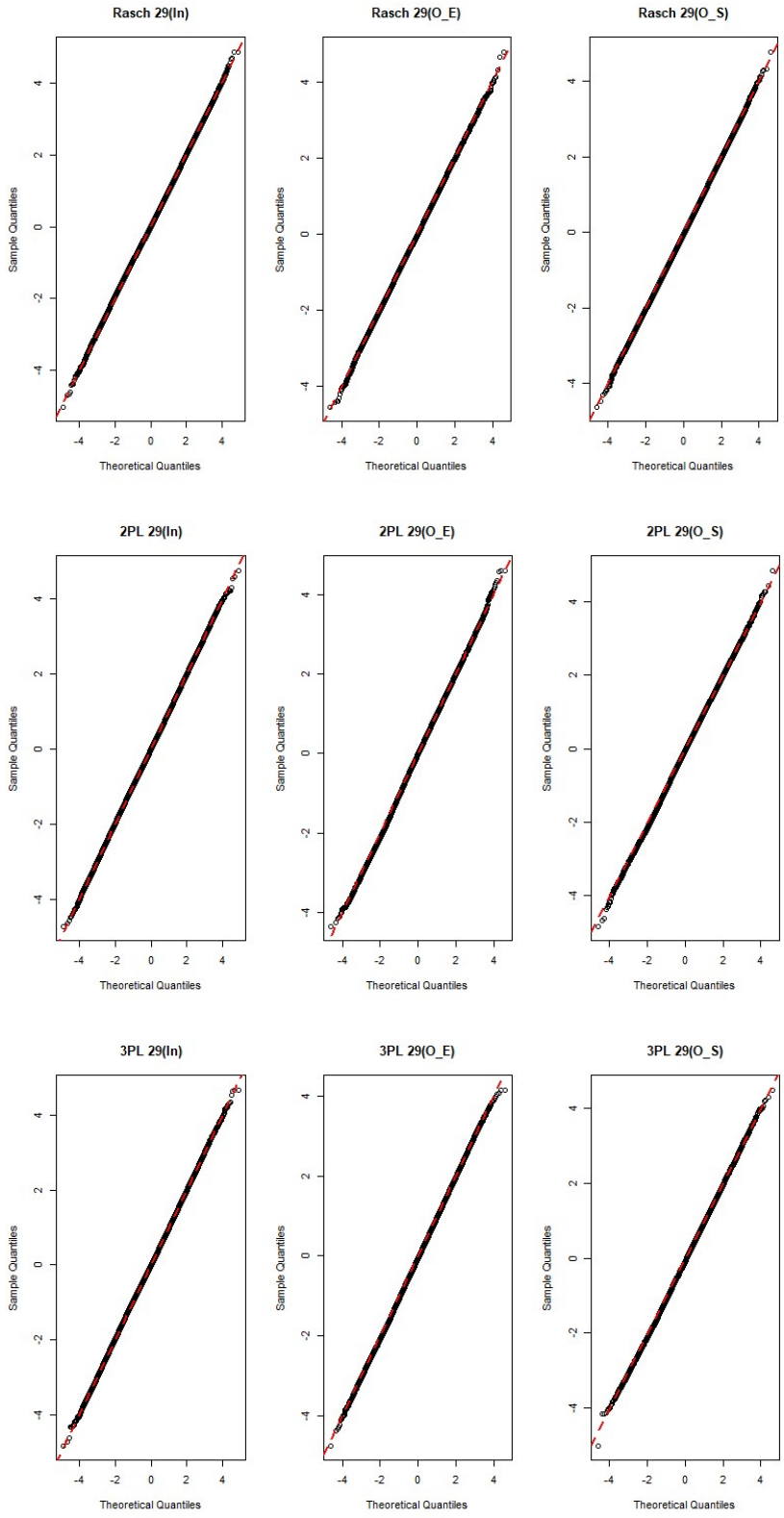


Figure C.28: RQR Checking Plot for Item 29.

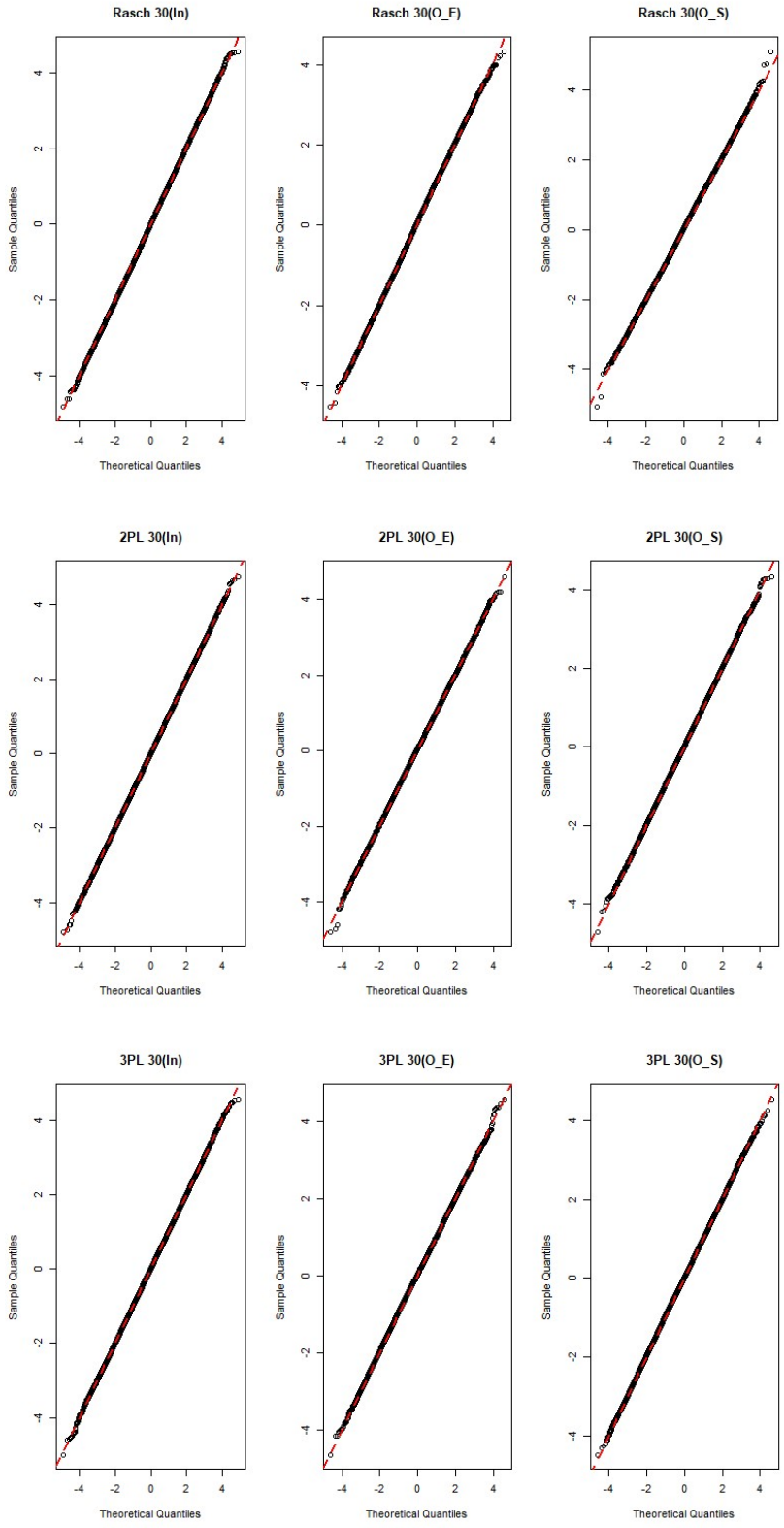


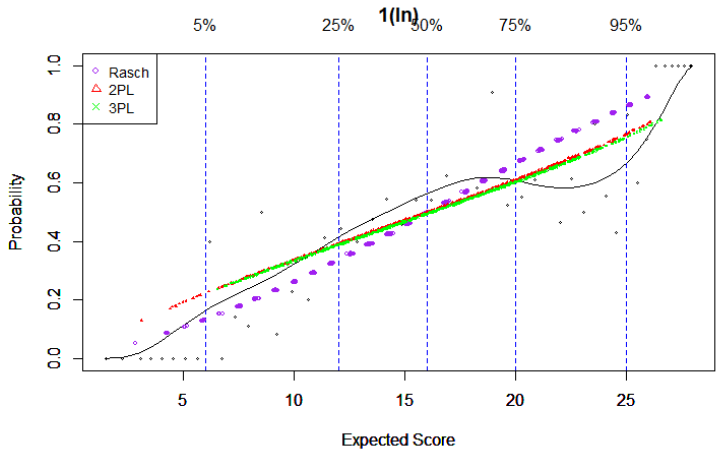
Figure C.29: RQR Checking Plot for Item 30.

Appendix D

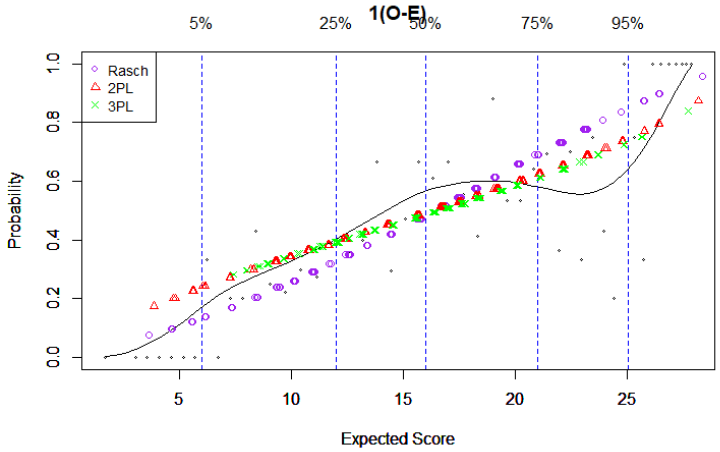
Kernel Smoothing Checking Plots

The Appendix D shows kernel smoothing checking plots for all items except for item 29 and the purpose is to convince the readers about our conclusion. All these plots are shown from next page. Each kernel smoothing checking plots is consisting of 3 small plots, which represents 3 cases. In kernel smoothing checking, the expected test takers' sum scores and their estimated success probabilities with respect to each item from Rasch, 2PL and 3PL models are put in the same kernel smoothing ICC plot for comparison. Here we mention that the kernel smoothing ICCs are a little bit different from parametric IRT model ICCs. The kernel smoothing is a non-parametric method which requires minimum model assumption, its ICCs are basically increasing, i.e., some parts of kernel smoothing ICCs are not strictly increasing (e.g., Figure D.5, Figure D.13).

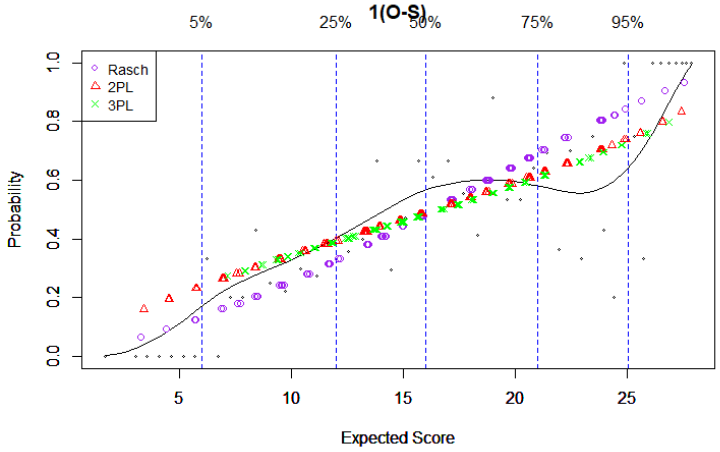
We use purple, red and green to denote Rasch, 2PL and 3PL model respectively. It can be seen that, when comparing to the kernel smoothing ICC, the estimated success probabilities from Rasch model show apparent deviation from estimated kernel smoothing ICC for many items (e.g., Figure D.1, Figure D.7, and Figure D.29), while there are no large deviations from kernel smoothing ICCs for those from both 2PL and 3PL model. Though the estimated success probabilities from 2PL and 3PL are close to each other for most of items, 2PL (true model) is closer to kernel smoothing ICCs than 3PL through some items, especially in the tail parts (e.g., Figure D.10, Figure D.11 and Figure D.23). So based on these checking plots, we can conclude that the kernel smoothing checking can serve as an effective way for IRT model assessment.



(a) Item 1: In-sample

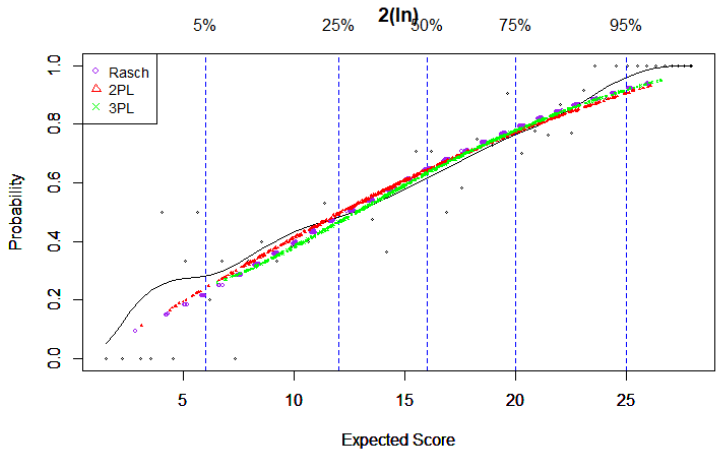


(b) Item 1: Out-of-sample-E

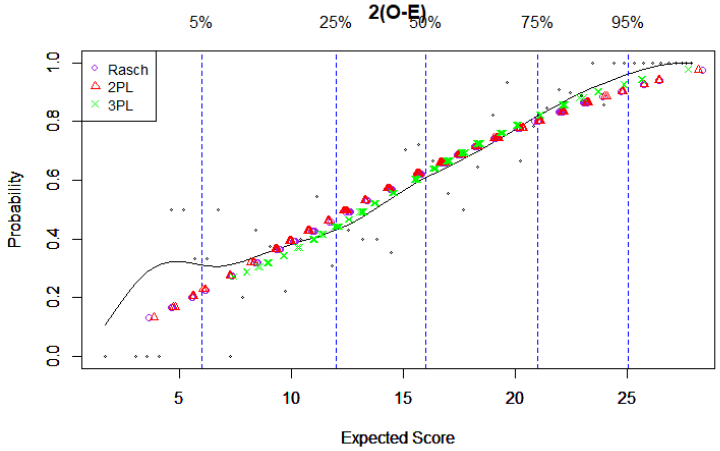


(c) Item 1: Out-of-sample-S

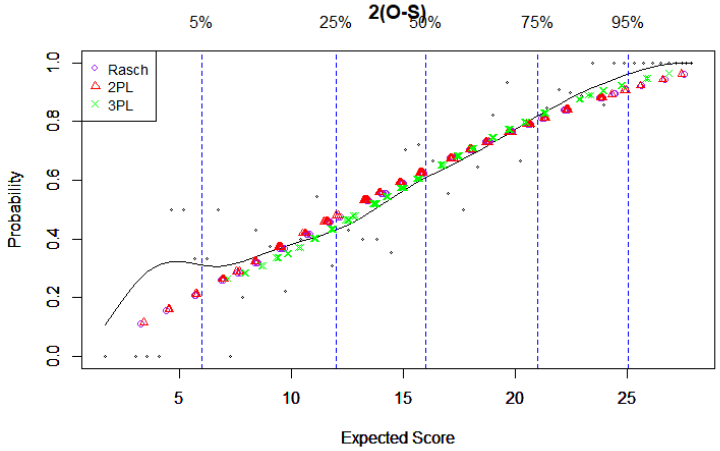
Figure D.1: Kernel Smoothing Checking Plot for Item 1.



(a) Item 2: In-sample

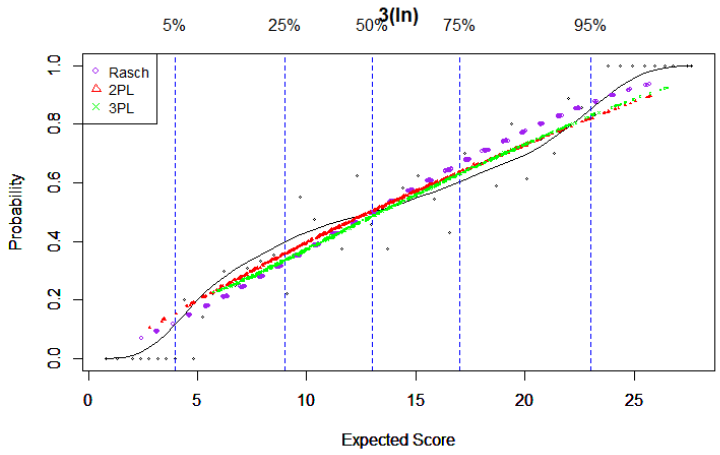


(b) Item 2: Out-of-sample-E

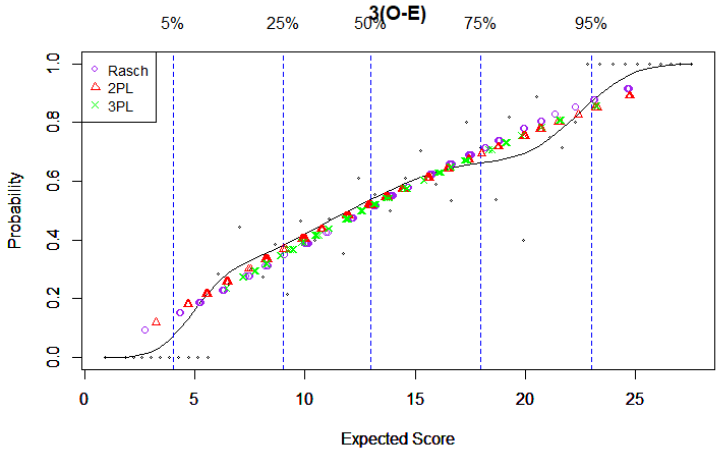


(c) Item 2: Out-of-sample-S

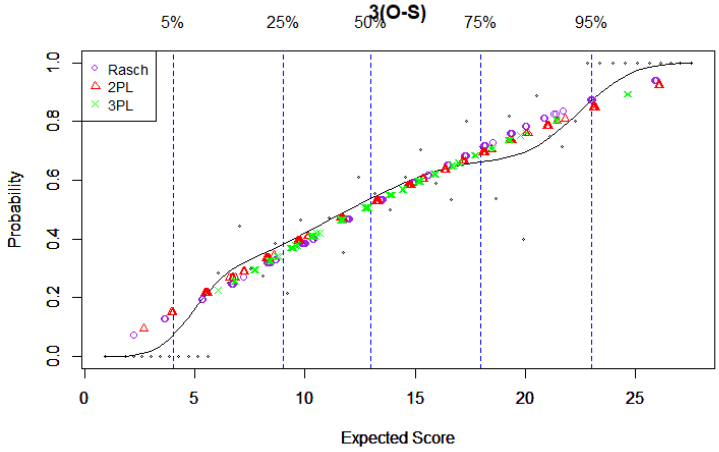
Figure D.2: Kernel Smoothing Checking Plot for Item 2.



(a) Item 3: In-sample

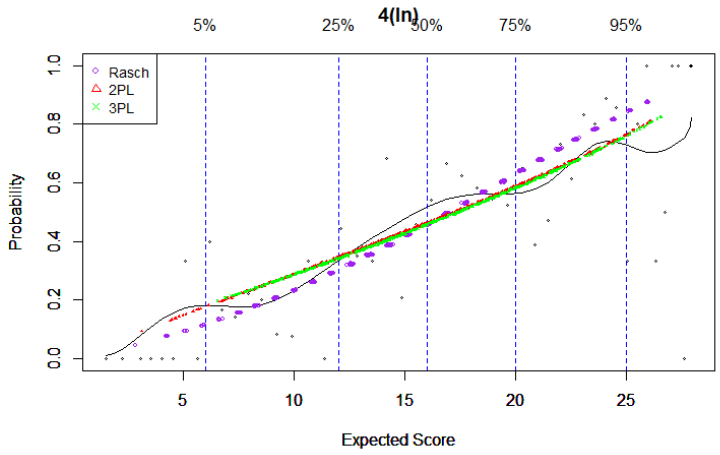


(b) Item 3: Out-of-sample-E

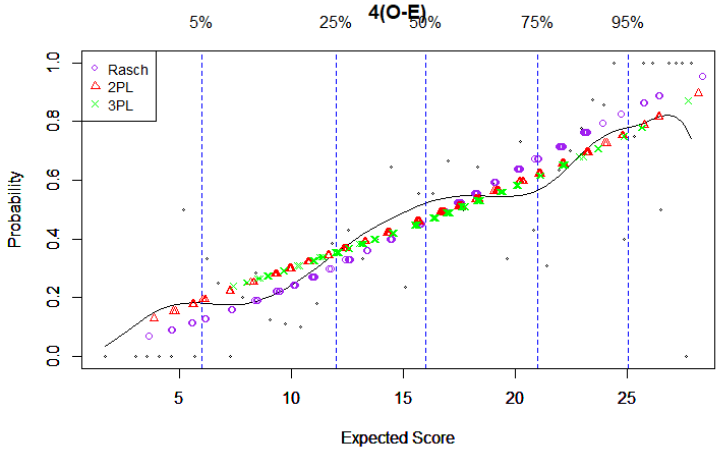


(c) Item 3: Out-of-sample-S

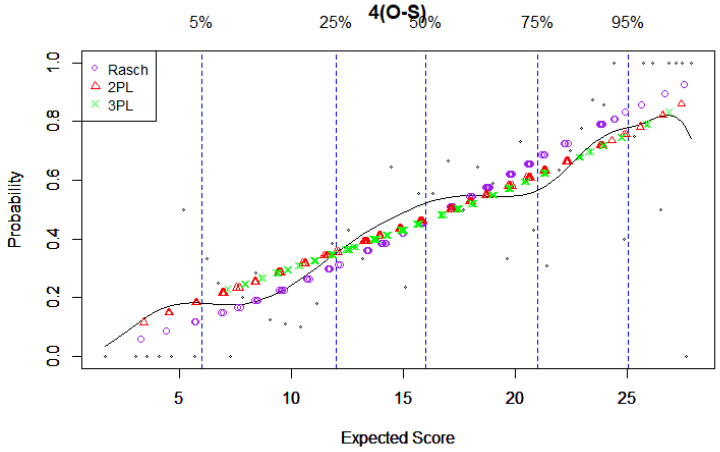
Figure D.3: Kernel Smoothing Checking Plot for Item 3.



(a) Item 4: In-sample

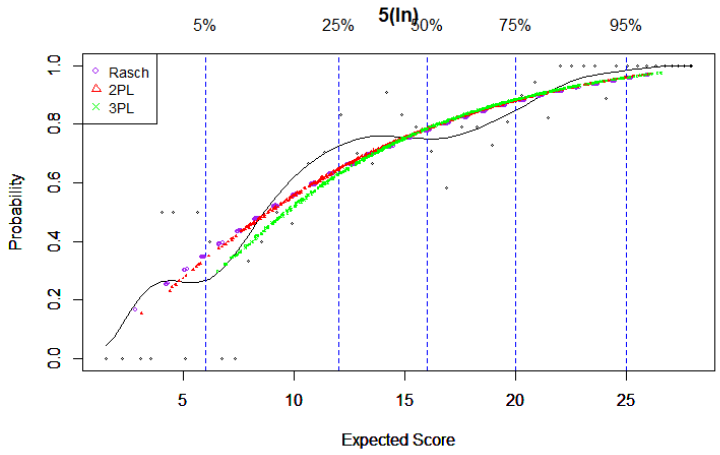


(b) Item 4: Out-of-sample-E

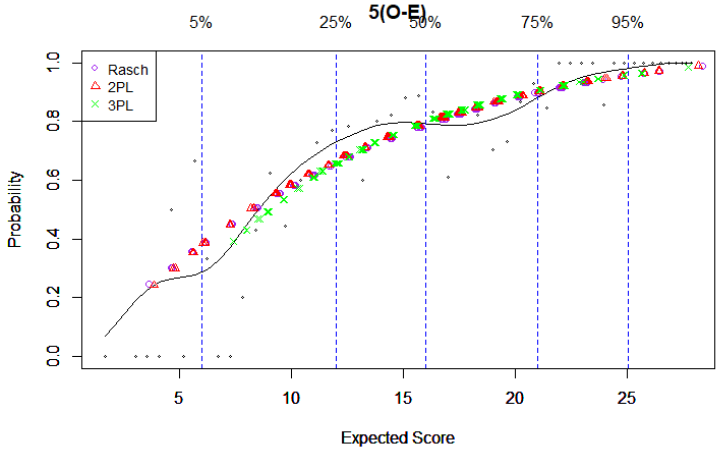


(c) Item 4: Out-of-sample-S

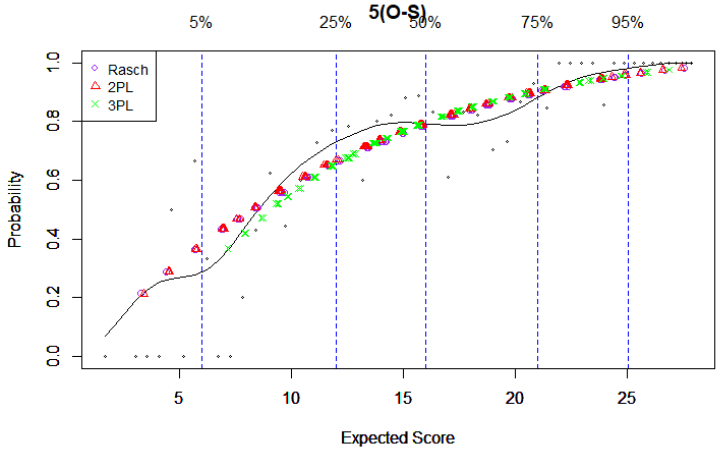
Figure D.4: Kernel Smoothing Checking Plot for Item 4.



(a) Item 5: In-sample

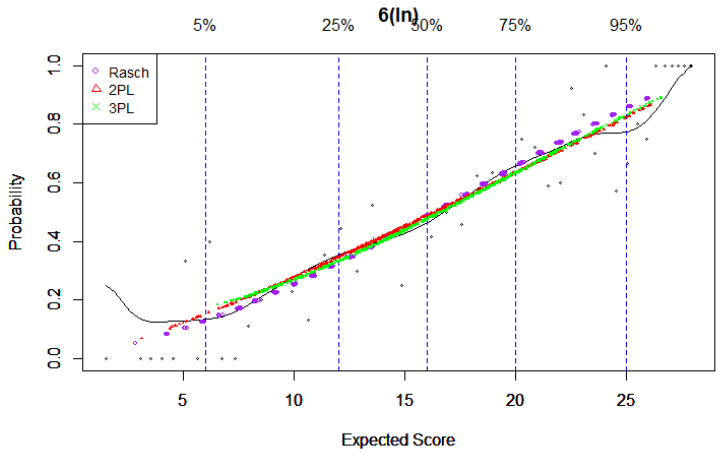


(b) Item 5: Out-of-sample-E

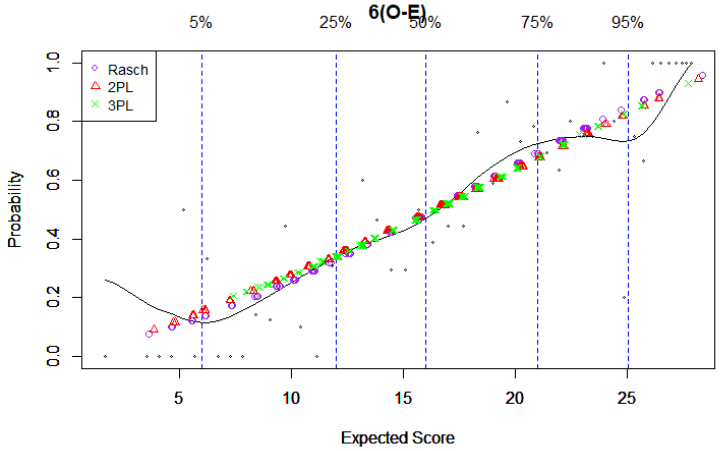


(c) Item 5: Out-of-sample-S

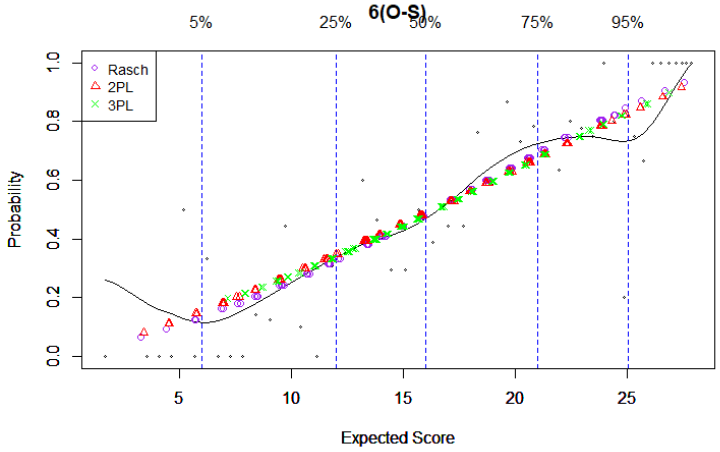
Figure D.5: Kernel Smoothing Checking Plot for Item 5.



(a) Item 6: In-sample

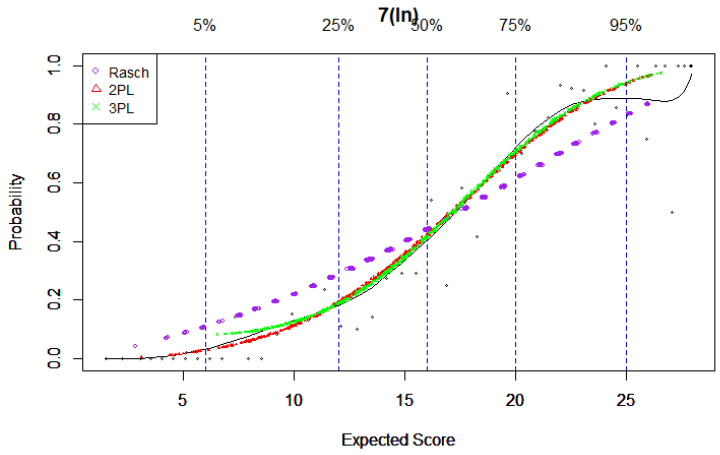


(b) Item 6: Out-of-sample-E

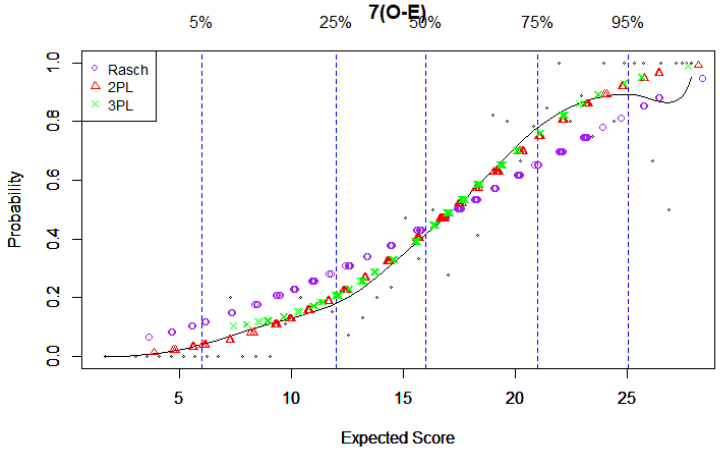


(c) Item 6: Out-of-sample-S

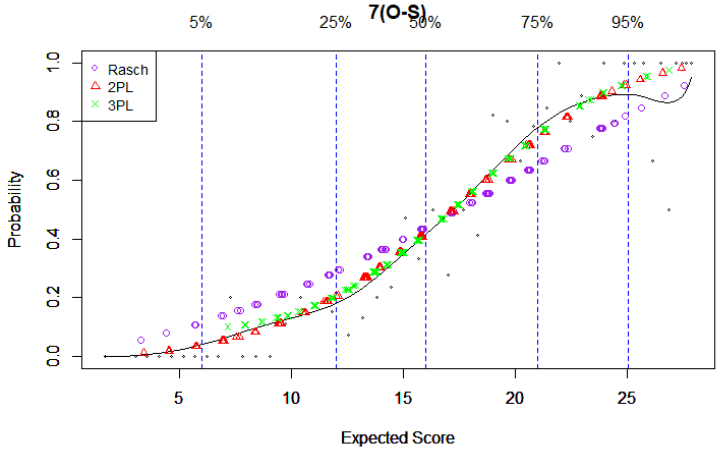
Figure D.6: Kernel Smoothing Checking Plot for Item 6.



(a) Item 7: In-sample

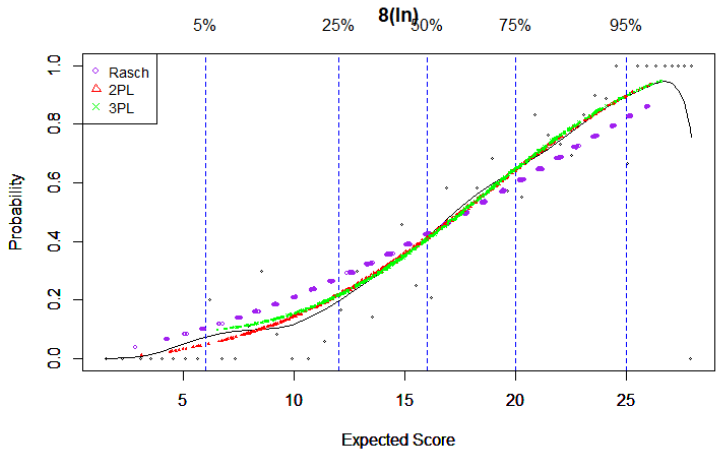


(b) Item 7: Out-of-sample-E

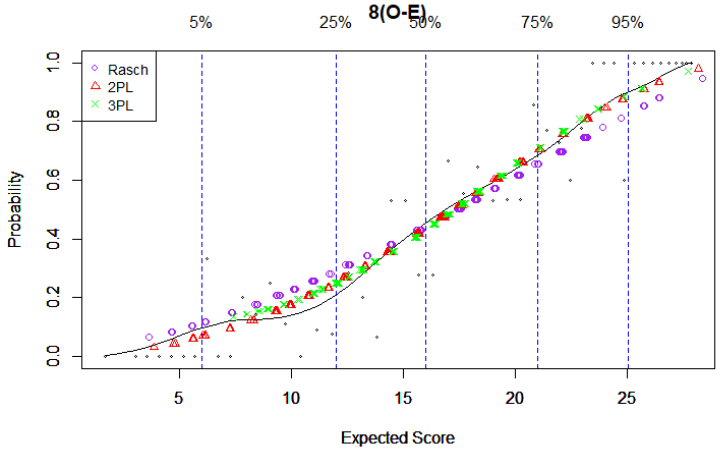


(c) Item 7: Out-of-sample-S

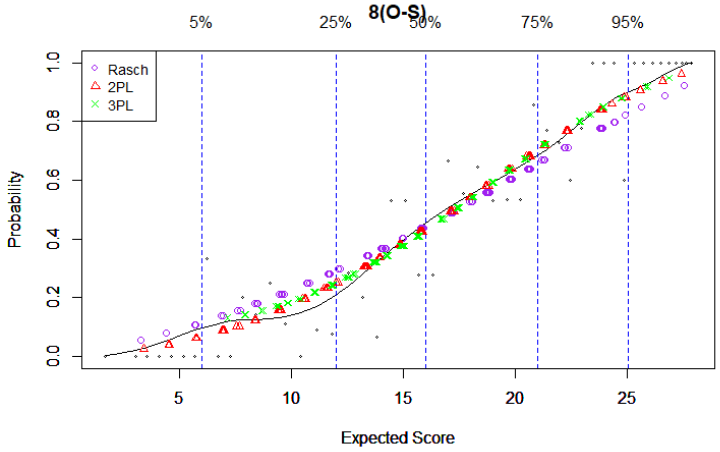
Figure D.7: Kernel Smoothing Checking Plot for Item 7.



(a) Item 8: In-sample

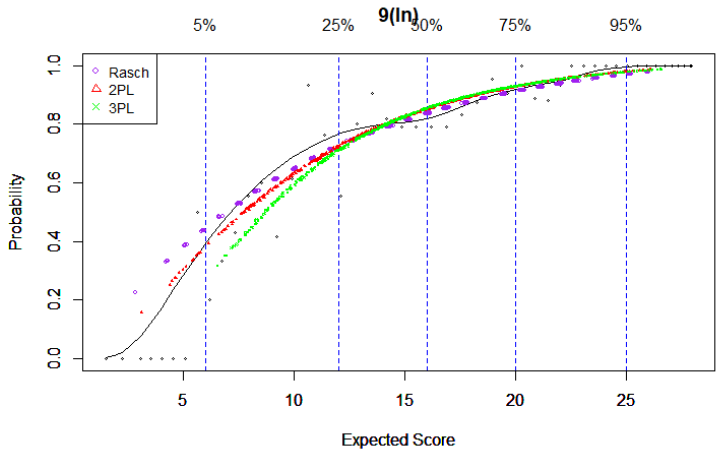


(b) Item 8: Out-of-sample-E

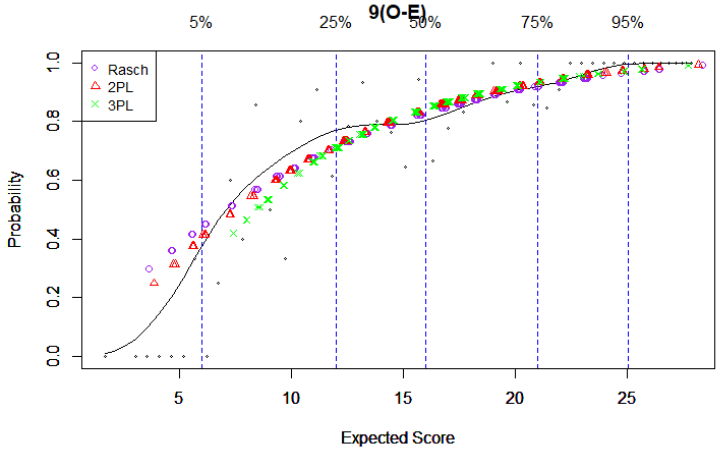


(c) Item 8: Out-of-sample-S

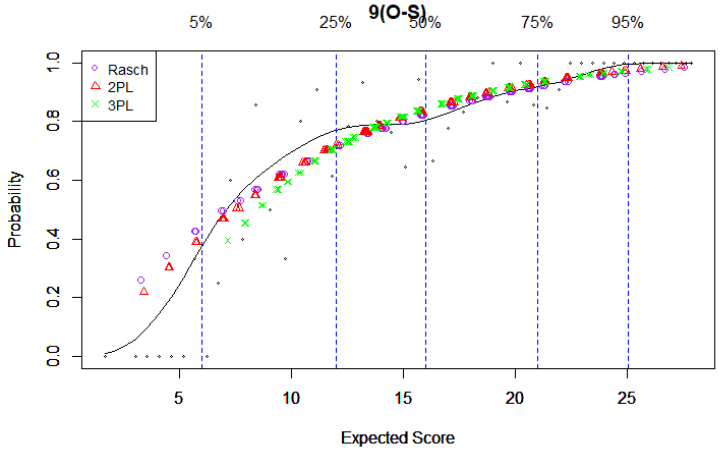
Figure D.8: Kernel Smoothing Checking Plot for Item 8.



(a) Item 9: In-sample

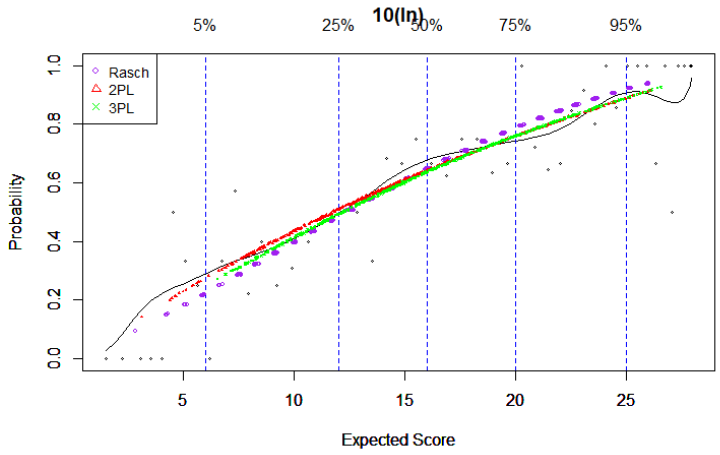


(b) Item 9: Out-of-sample-E

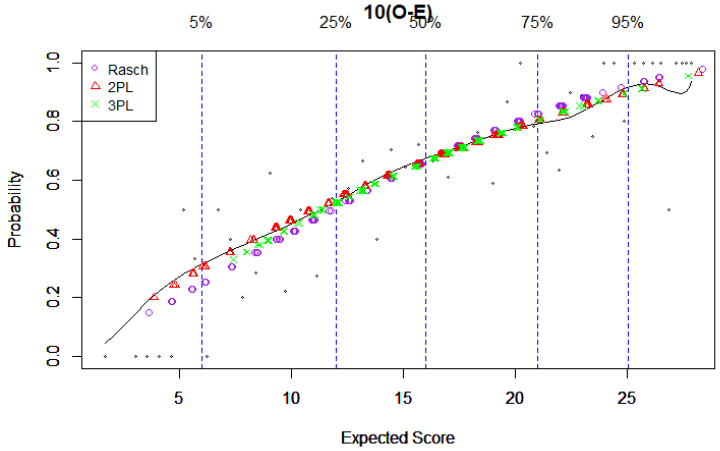


(c) Item 9: Out-of-sample-S

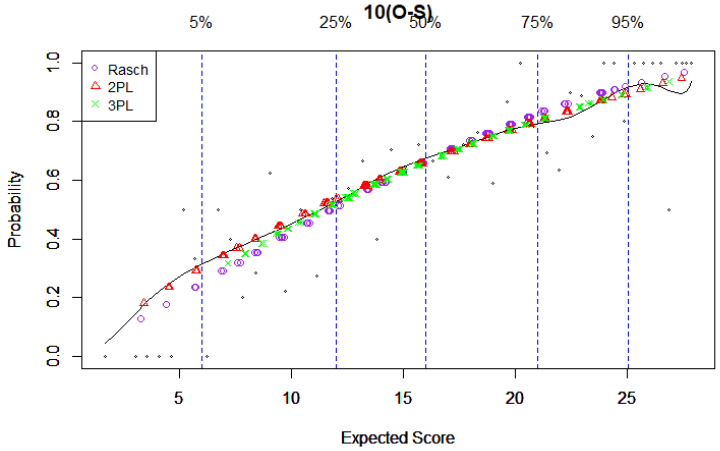
Figure D.9: Kernel Smoothing Checking Plot for Item 9.



(a) Item 10: In-sample

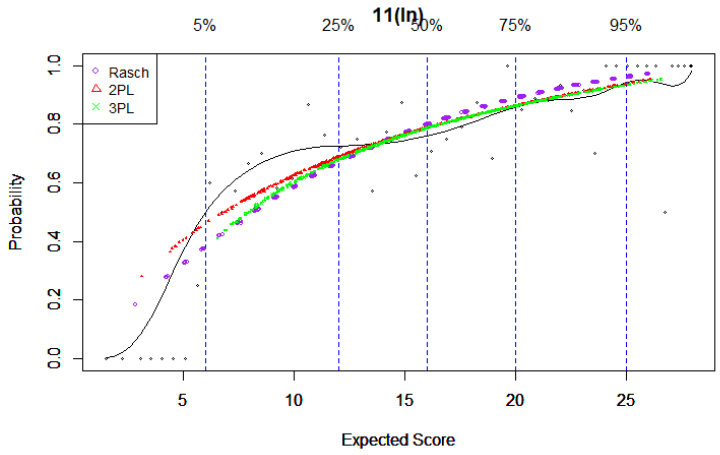


(b) Item 10: Out-of-sample-E

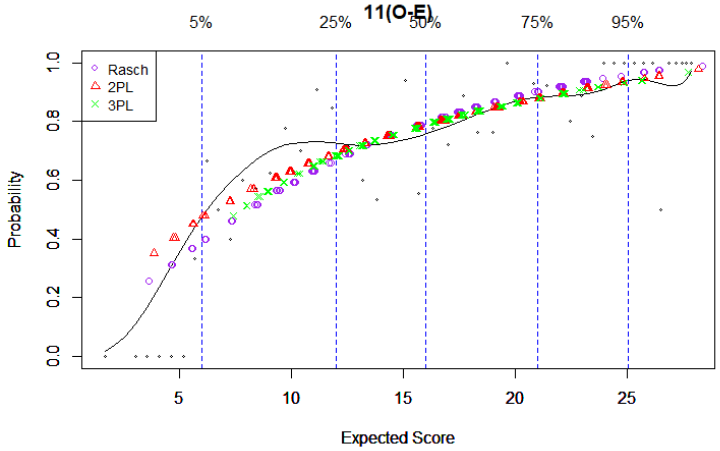


(c) Item 10: Out-of-sample-S

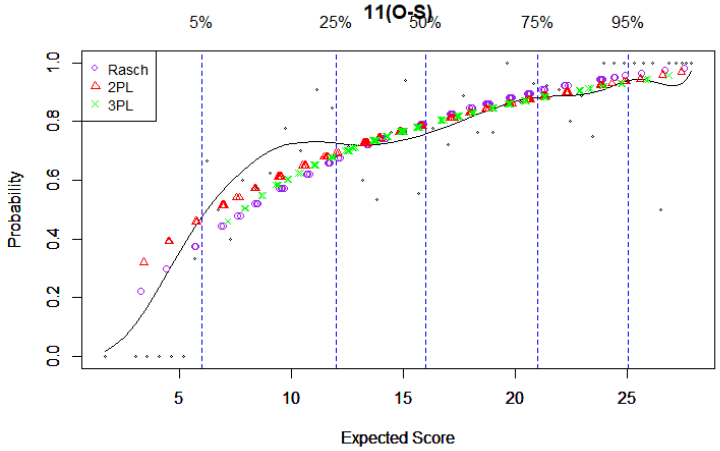
Figure D.10: Kernel Smoothing Checking Plot for Item 10.



(a) Item 11: In-sample

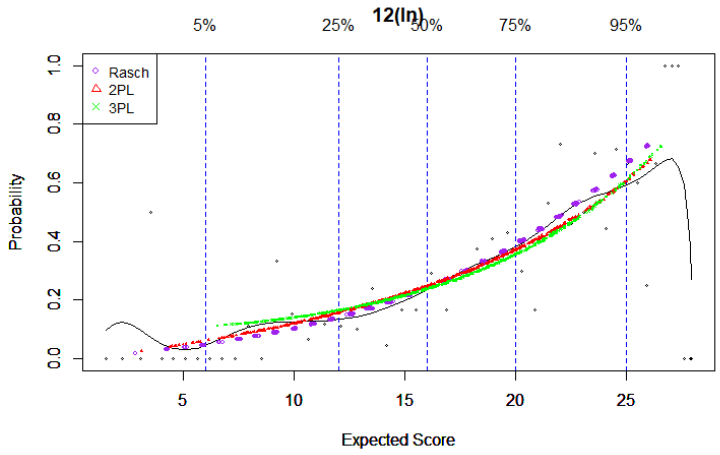


(b) Item 11: Out-of-sample-E

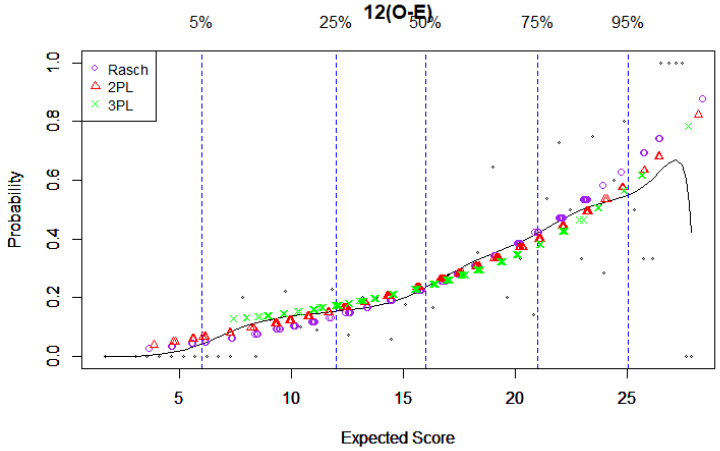


(c) Item 11: Out-of-sample-S

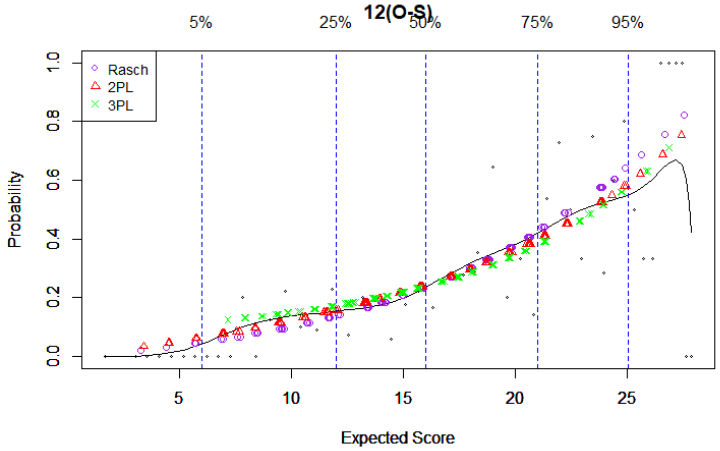
Figure D.11: Kernel Smoothing Checking Plot for Item 11.



(a) Item 12: In-sample

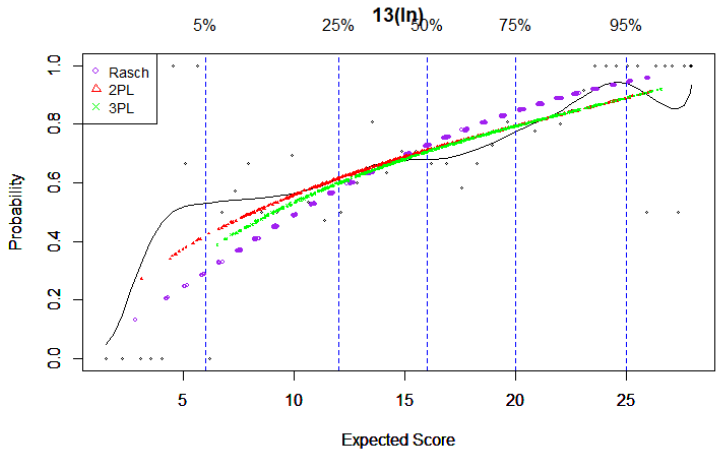


(b) Item 12: Out-of-sample-E

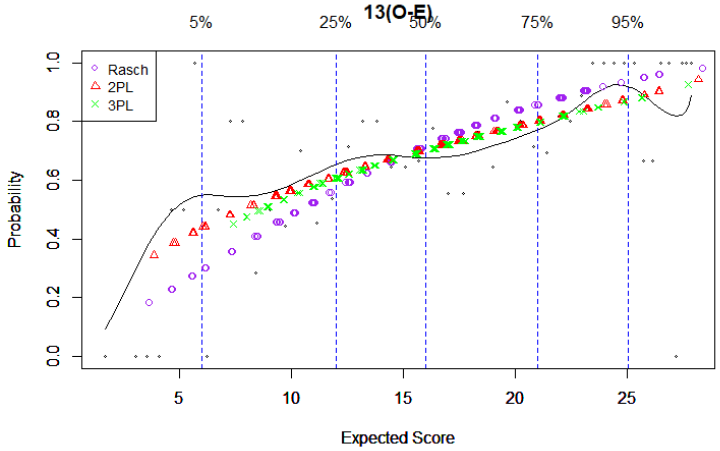


(c) Item 12: Out-of-sample-S

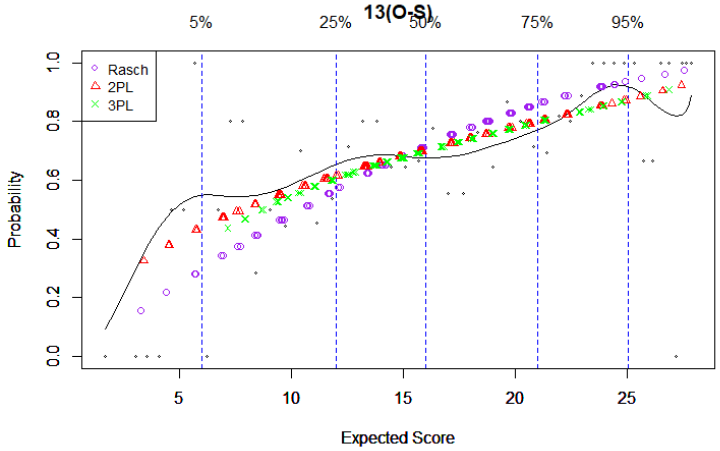
Figure D.12: Kernel Smoothing Checking Plot for Item 12.



(a) Item 13: In-sample

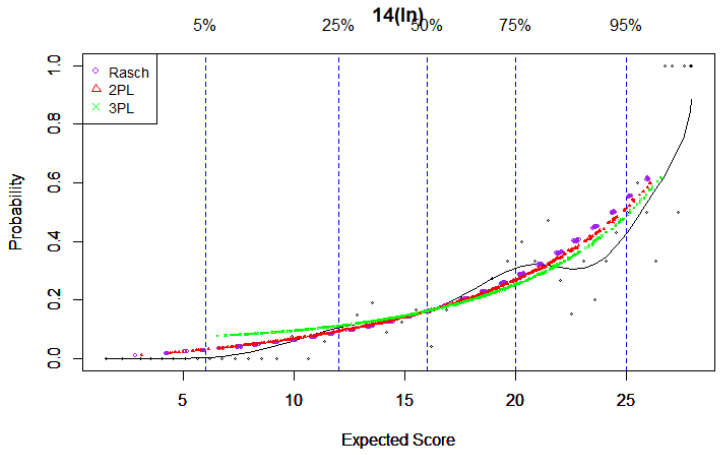


(b) Item 13: Out-of-sample-E

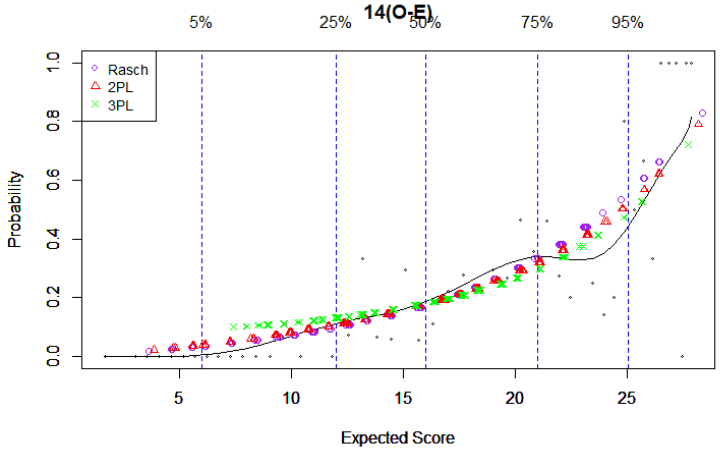


(c) Item 13: Out-of-sample-S

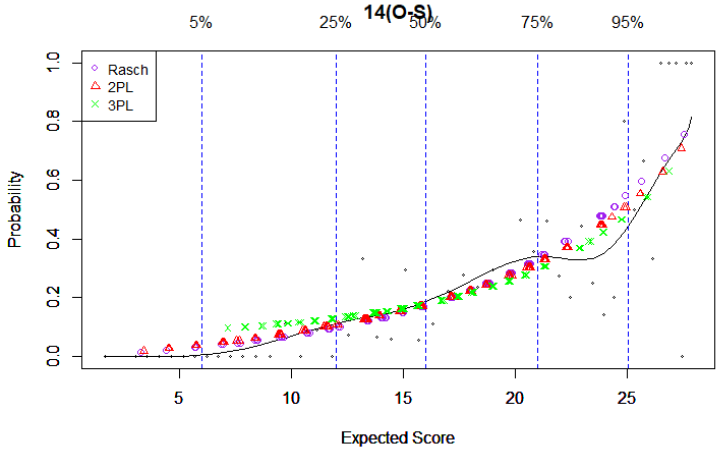
Figure D.13: Kernel Smoothing Checking Plot for Item 13.



(a) Item 14: In-sample

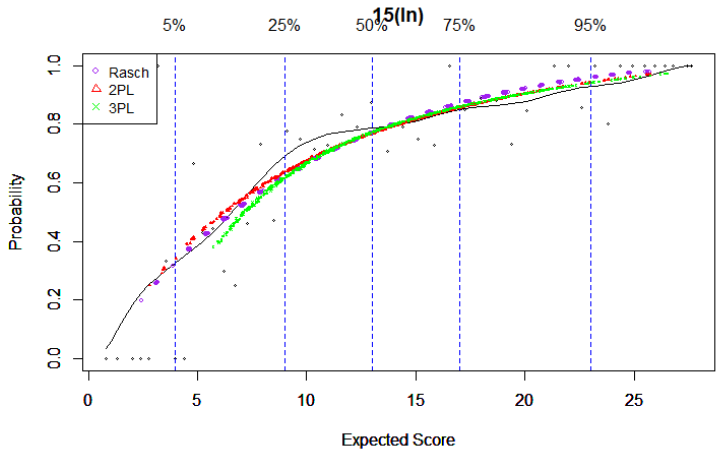


(b) Item 14: Out-of-sample-E

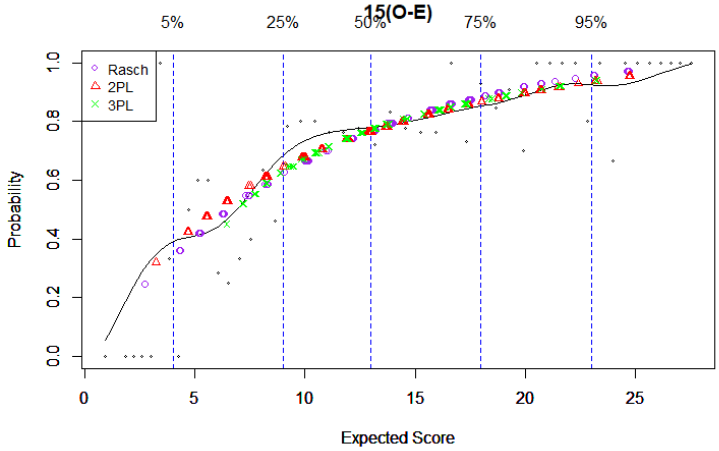


(c) Item 14: Out-of-sample-S

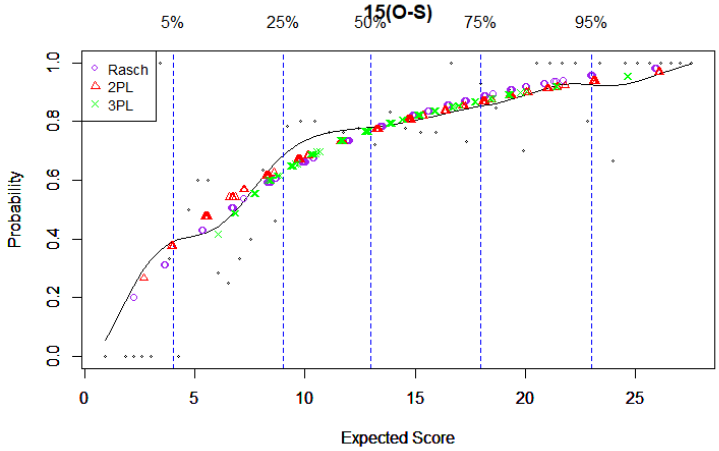
Figure D.14: Kernel Smoothing Checking Plot for Item 14.



(a) Item 15: In-sample

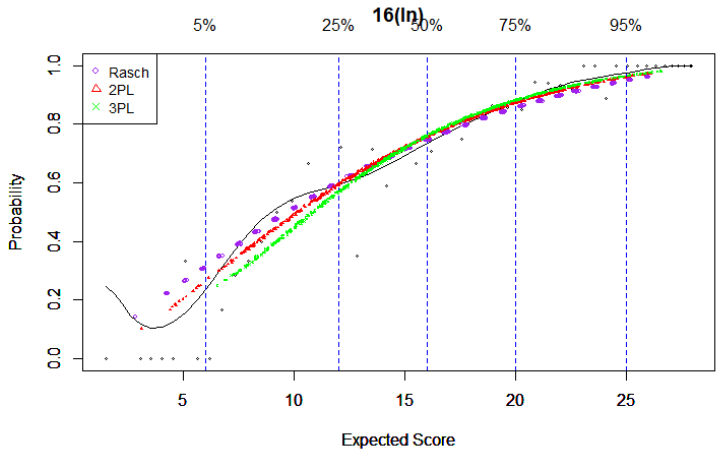


(b) Item 15: Out-of-sample-E

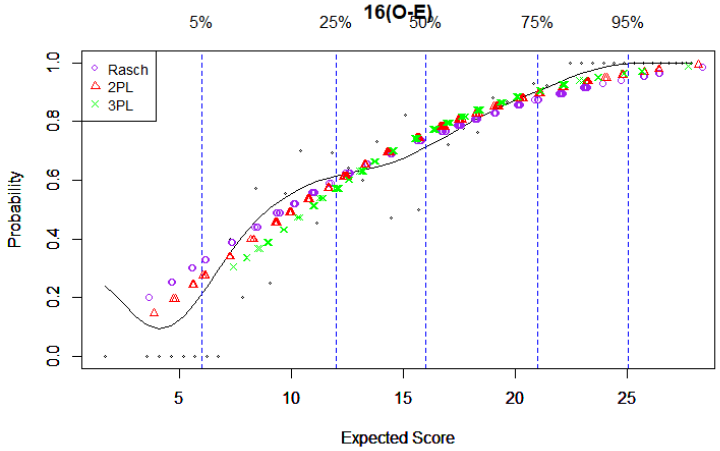


(c) Item 15: Out-of-sample-S

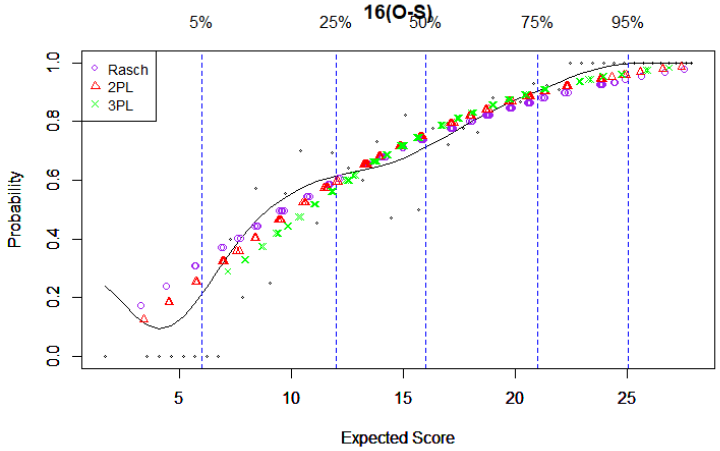
Figure D.15: Kernel Smoothing Checking Plot for Item 15.



(a) Item 16: In-sample

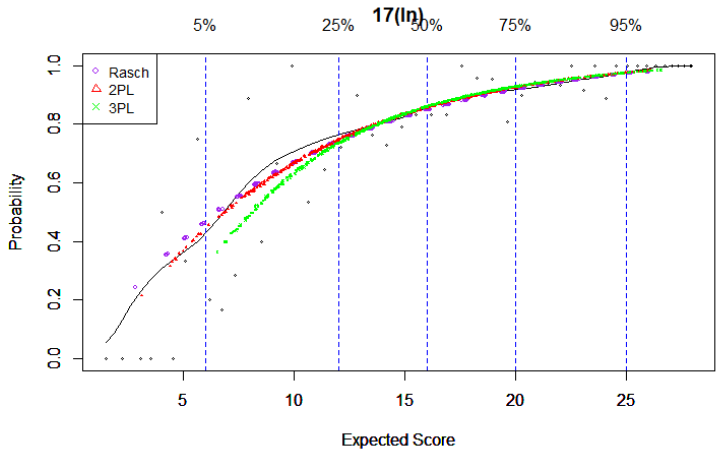


(b) Item 16: Out-of-sample-E

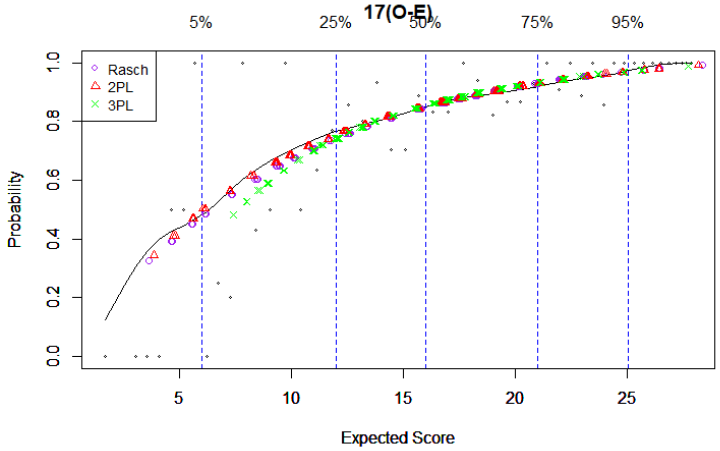


(c) Item 16: Out-of-sample-S

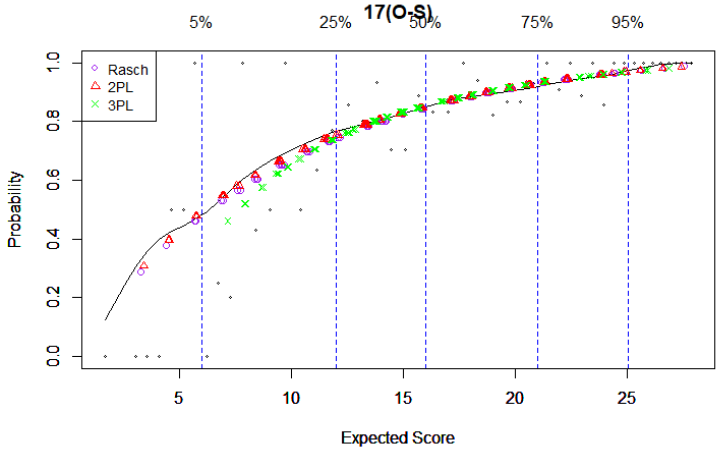
Figure D.16: Kernel Smoothing Checking Plot for Item 16.



(a) Item 17: In-sample

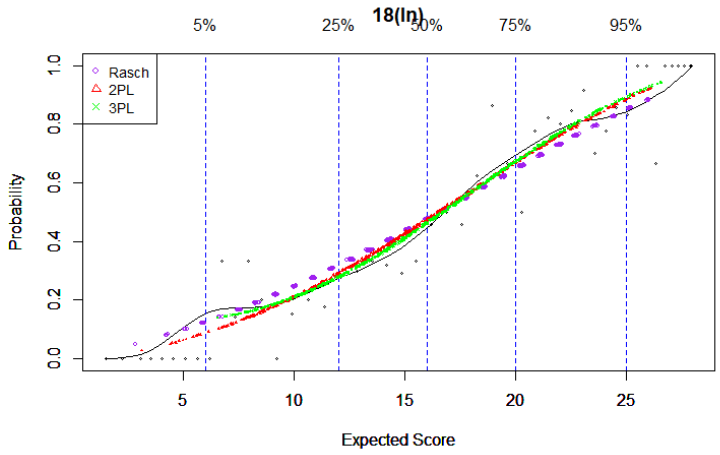


(b) Item 17: Out-of-sample-E

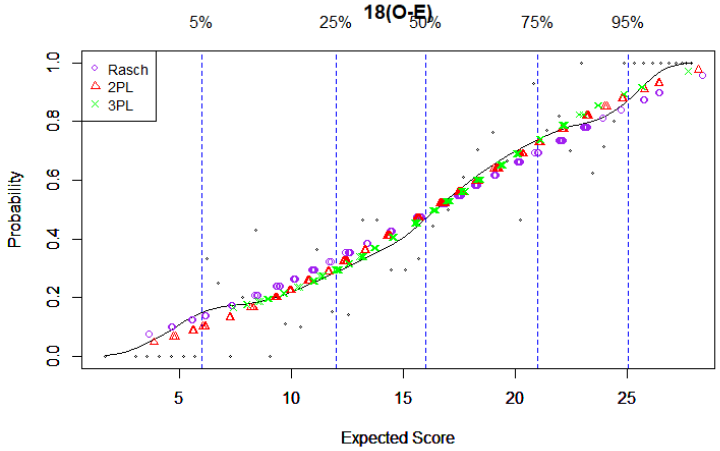


(c) Item 17: Out-of-sample-S

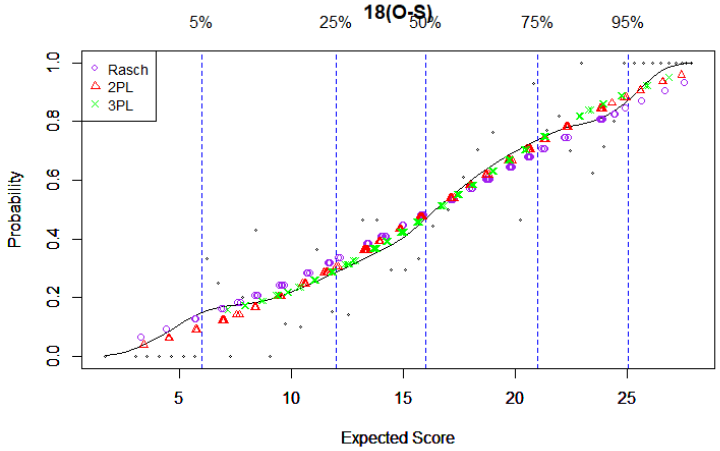
Figure D.17: Kernel Smoothing Checking Plot for Item 17.



(a) Item 18: In-sample

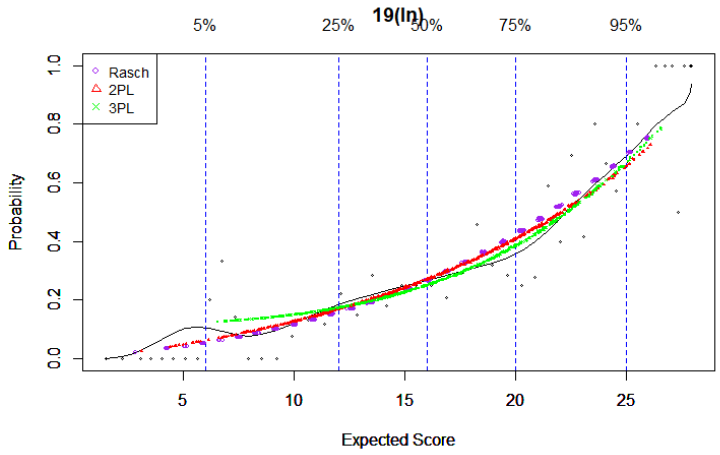


(b) Item 18: Out-of-sample-E

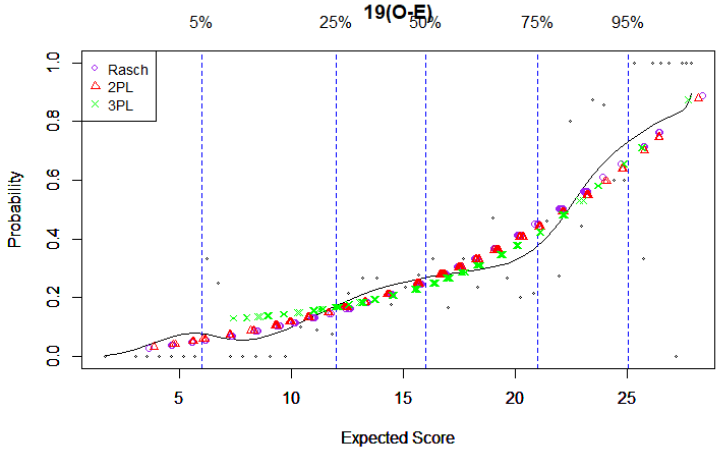


(c) Item 18: Out-of-sample-S

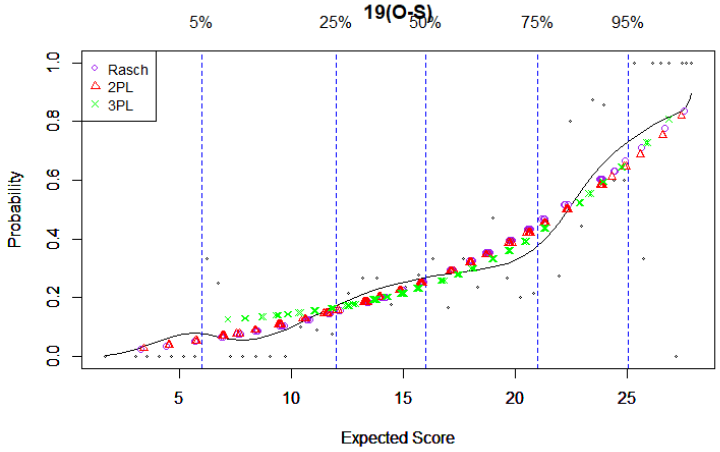
Figure D.18: Kernel Smoothing Checking Plot for Item 18.



(a) Item 19: In-sample

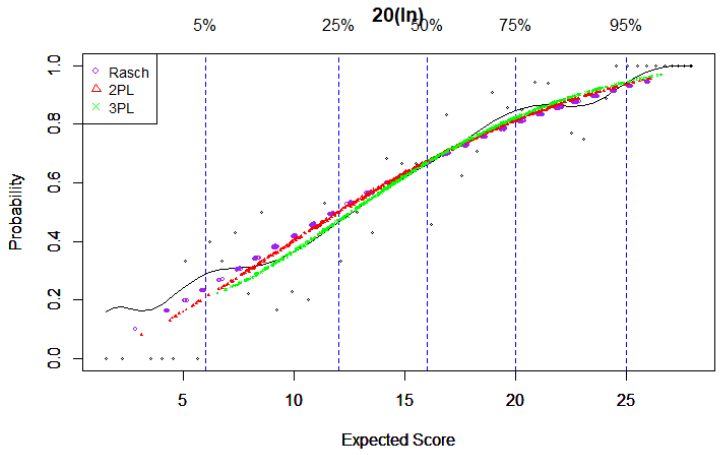


(b) Item 19: Out-of-sample-E

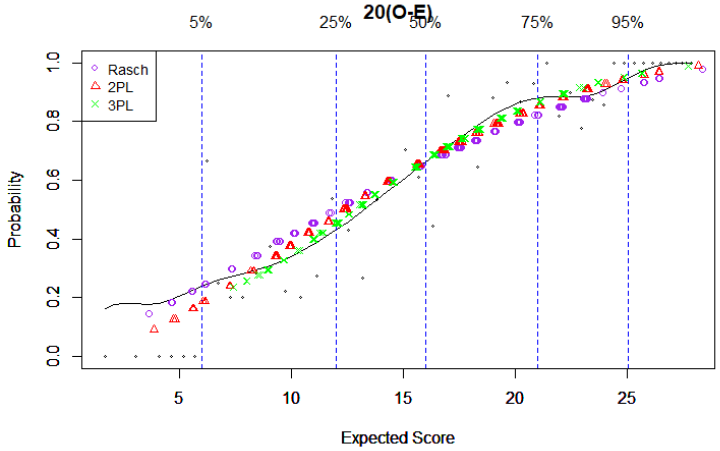


(c) Item 19: Out-of-sample-S

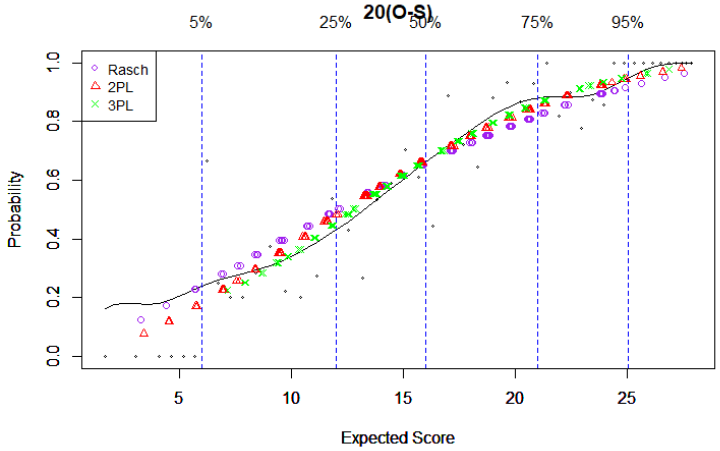
Figure D.19: Kernel Smoothing Checking Plot for Item 19.



(a) Item 20: In-sample

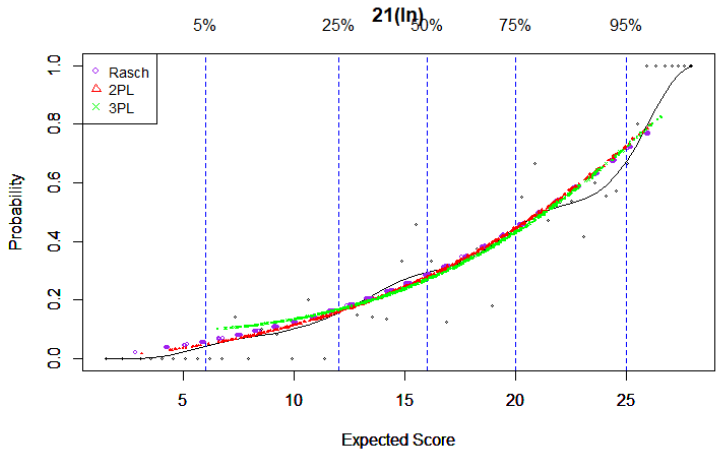


(b) Item 20: Out-of-sample-E

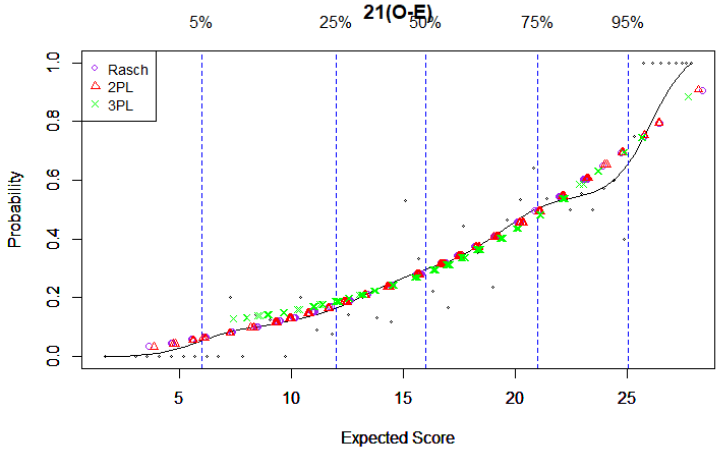


(c) Item 20: Out-of-sample-S

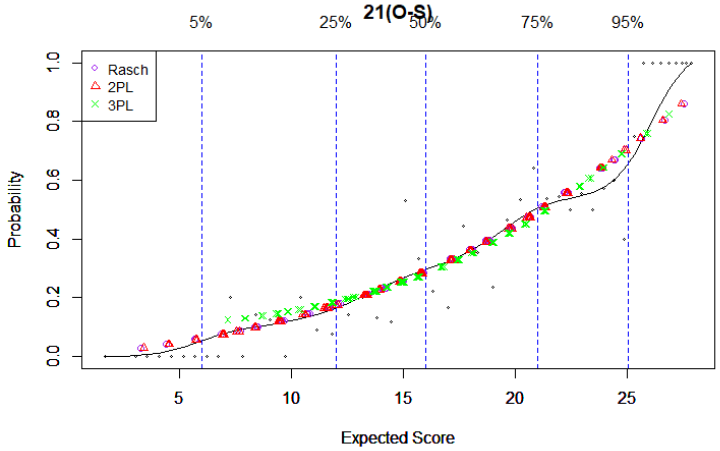
Figure D.20: Kernel Smoothing Checking Plot for Item 20.



(a) Item 21: In-sample

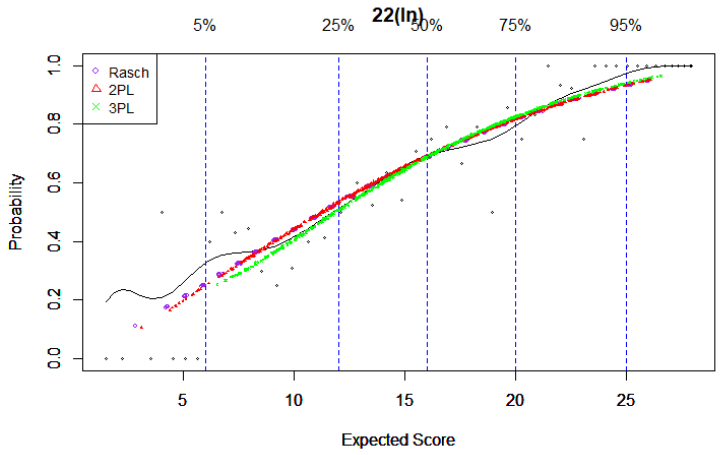


(b) Item 21: Out-of-sample-E

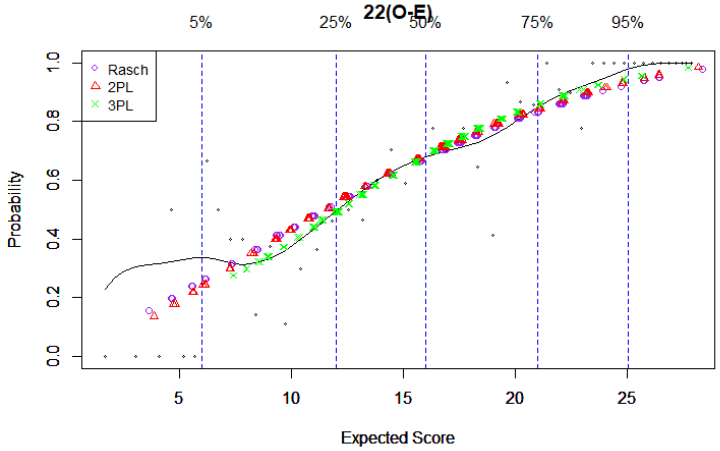


(c) Item 21: Out-of-sample-S

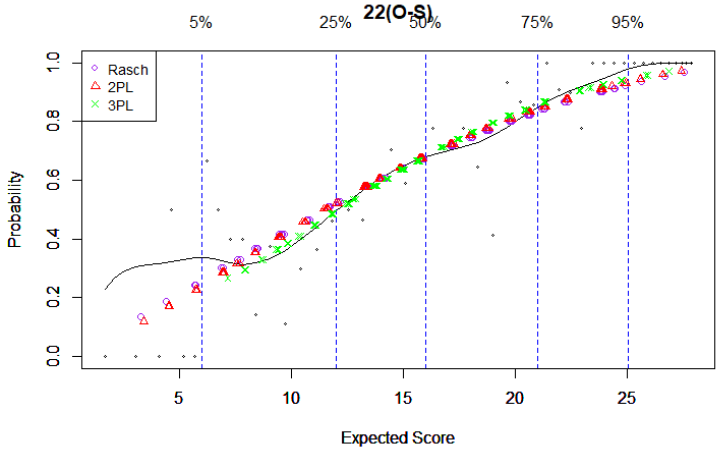
Figure D.21: Kernel Smoothing Checking Plot for Item 21.



(a) Item 22: In-sample

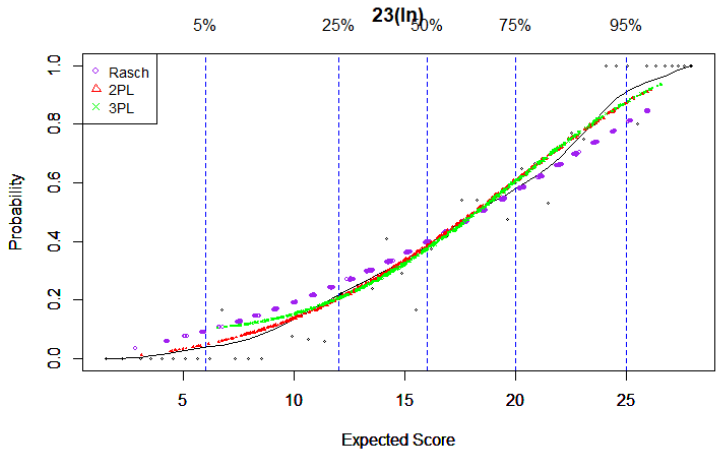


(b) Item 22: Out-of-sample-E

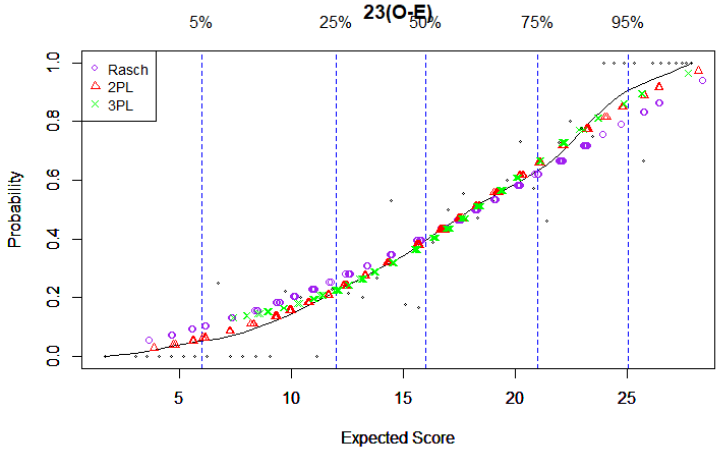


(c) Item 22: Out-of-sample-S

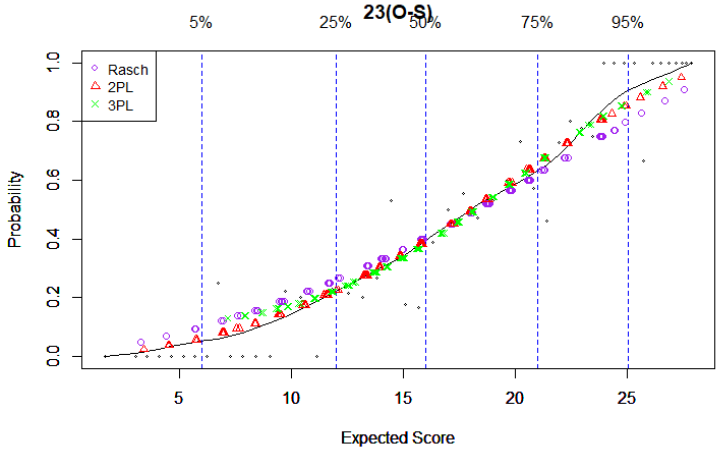
Figure D.22: Kernel Smoothing Checking Plot for Item 22.



(a) Item 23: In-sample

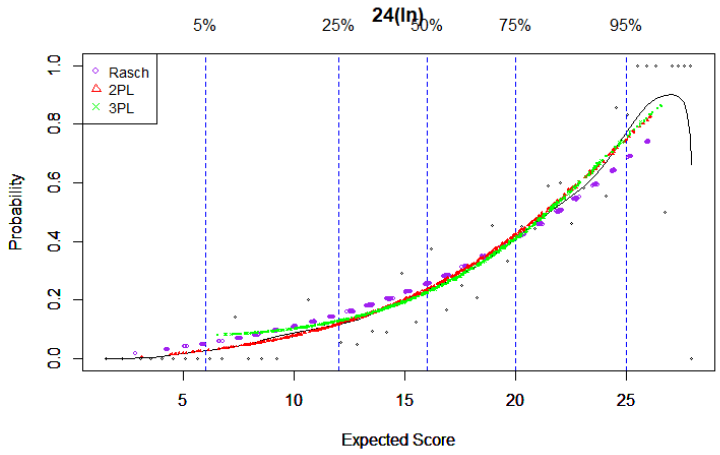


(b) Item 23: Out-of-sample-E

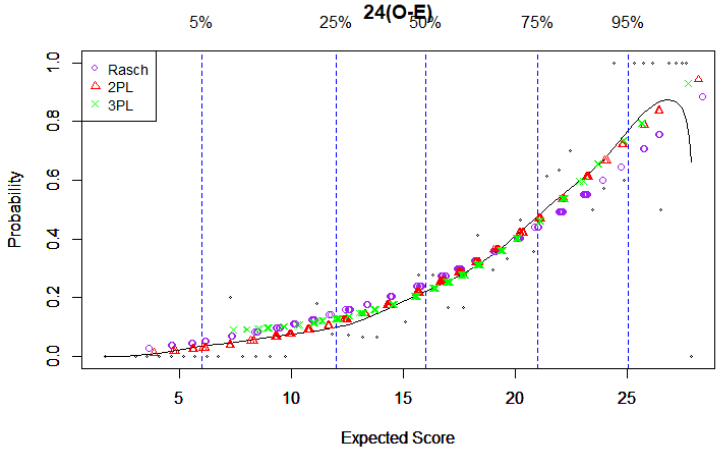


(c) Item 23: Out-of-sample-S

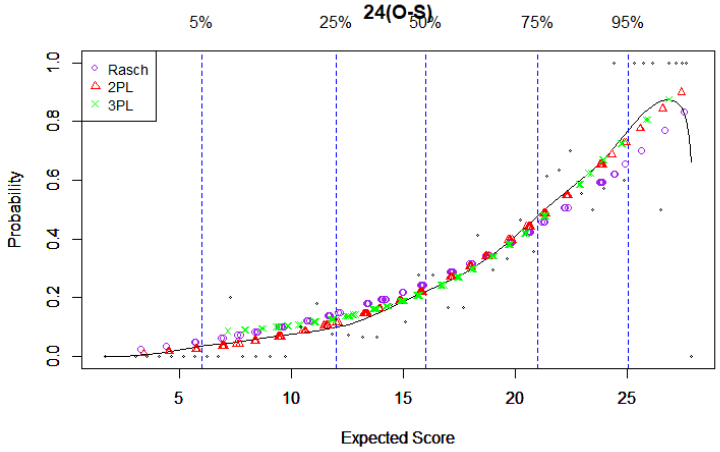
Figure D.23: Kernel Smoothing Checking Plot for Item 23.



(a) Item 24: In-sample

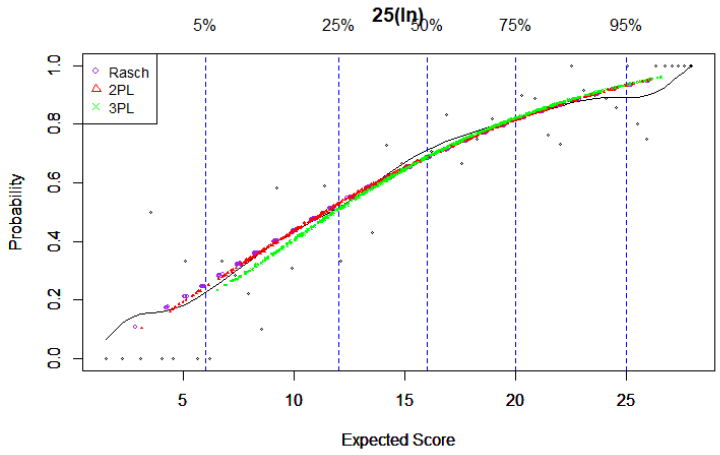


(b) Item 24: Out-of-sample-E

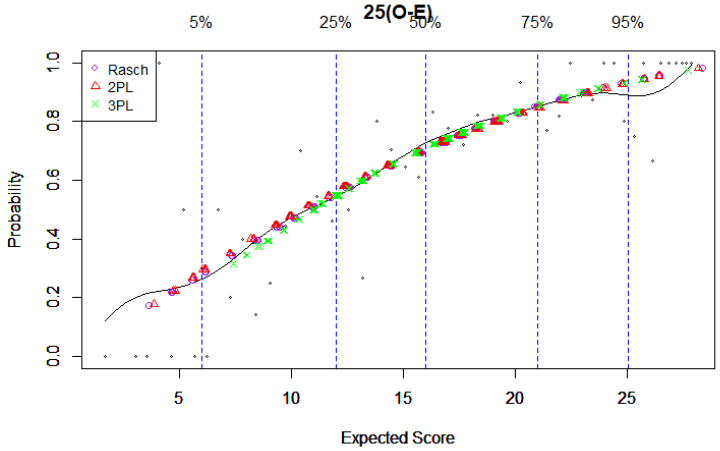


(c) Item 24: Out-of-sample-S

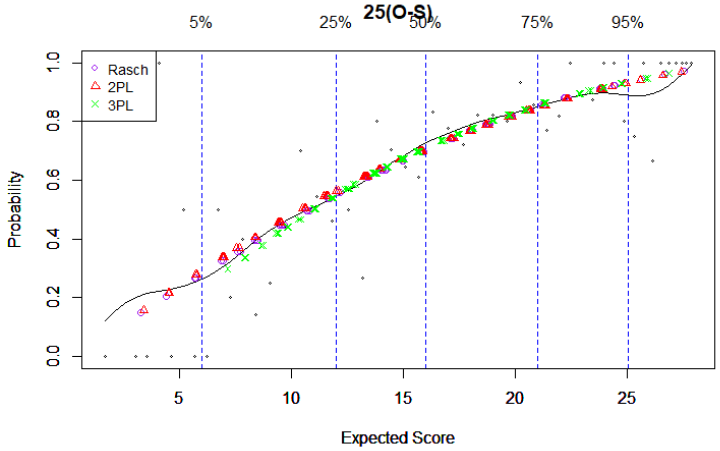
Figure D.24: Kernel Smoothing Checking Plot for Item 24.



(a) Item 25: In-sample

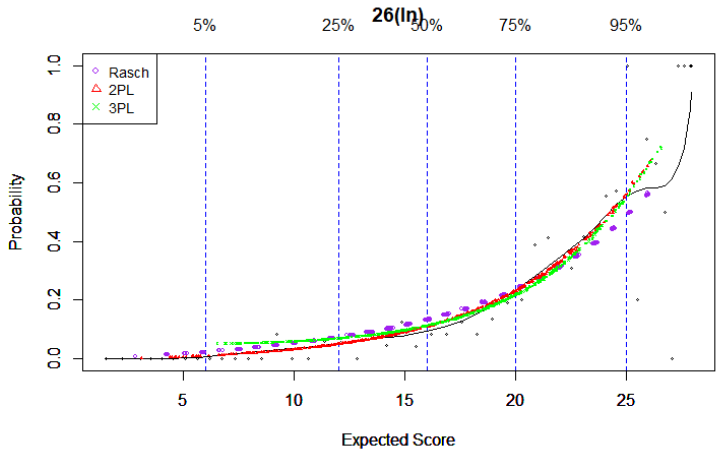


(b) Item 25: Out-of-sample-E

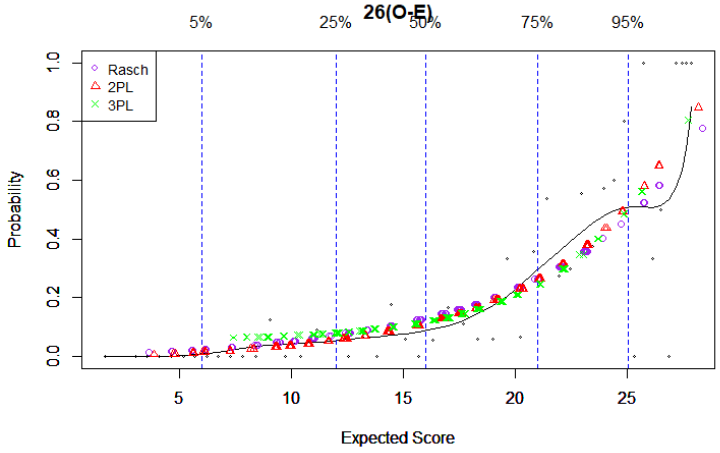


(c) Item 25: Out-of-sample-S

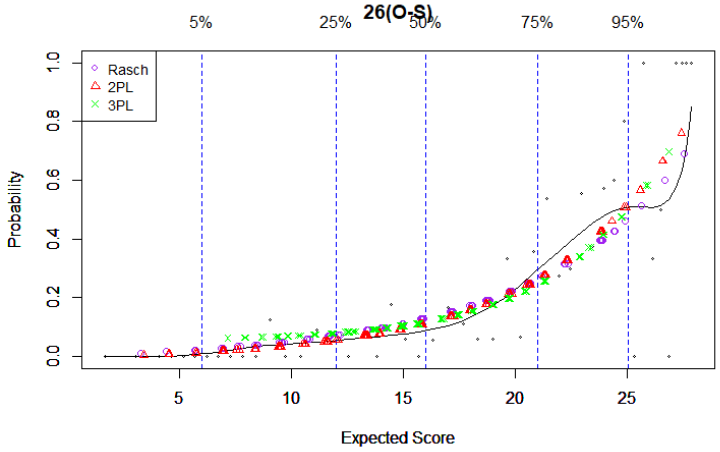
Figure D.25: Kernel Smoothing Checking Plot for Item 25.



(a) Item 26: In-sample

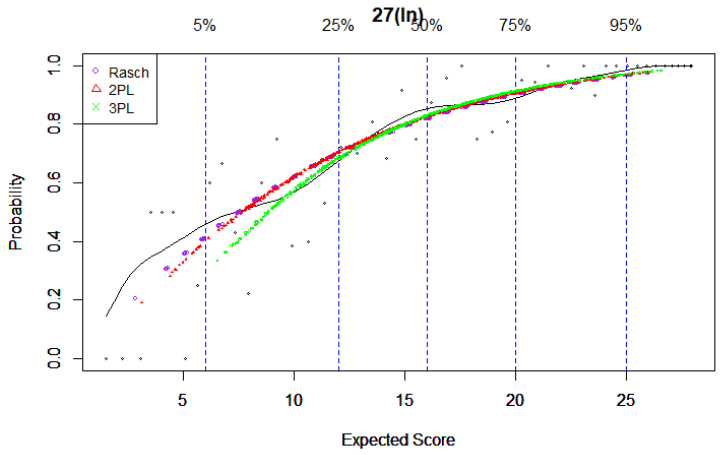


(b) Item 26: Out-of-sample-E

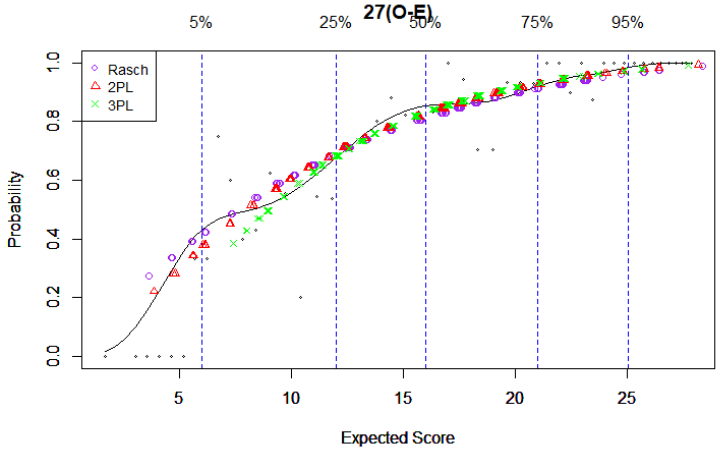


(c) Item 26: Out-of-sample-S

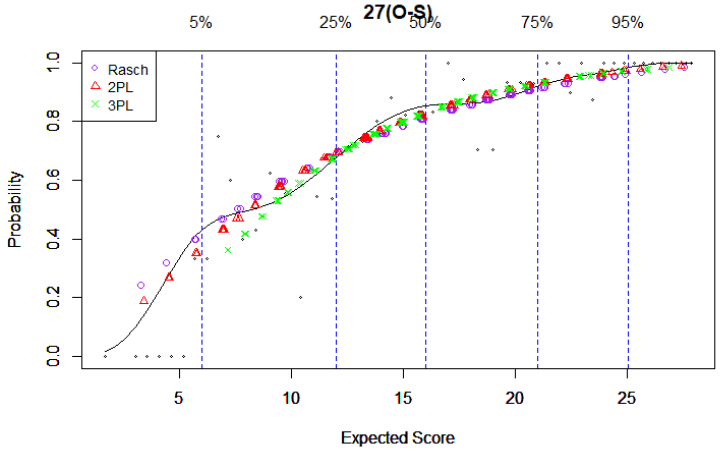
Figure D.26: Kernel Smoothing Checking Plot for Item 26.



(a) Item 27: In-sample

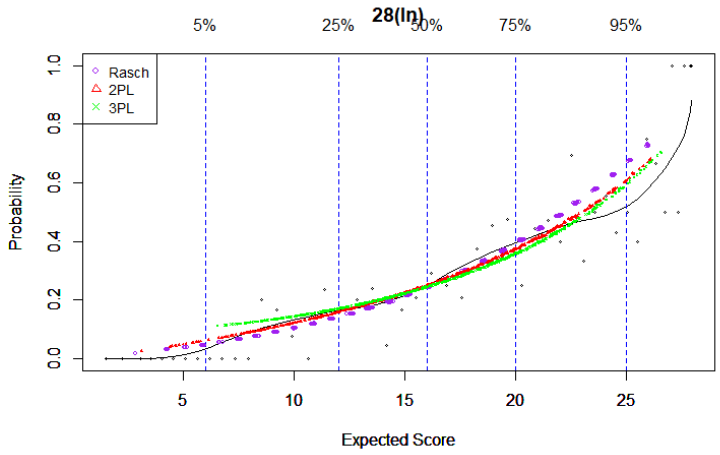


(b) Item 27: Out-of-sample-E

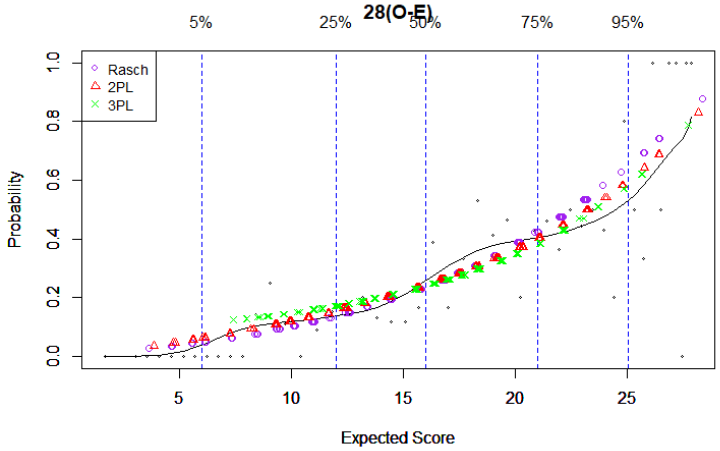


(c) Item 27: Out-of-sample-S

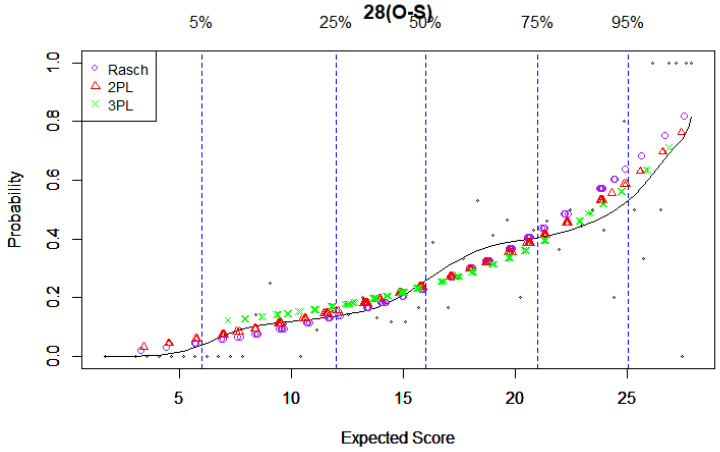
Figure D.27: Kernel Smoothing Checking Plot for Item 27.



(a) Item 28: In-sample

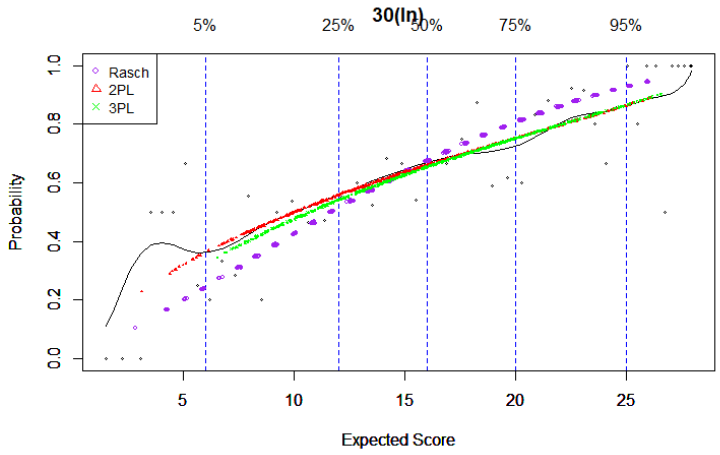


(b) Item 28: Out-of-sample-E

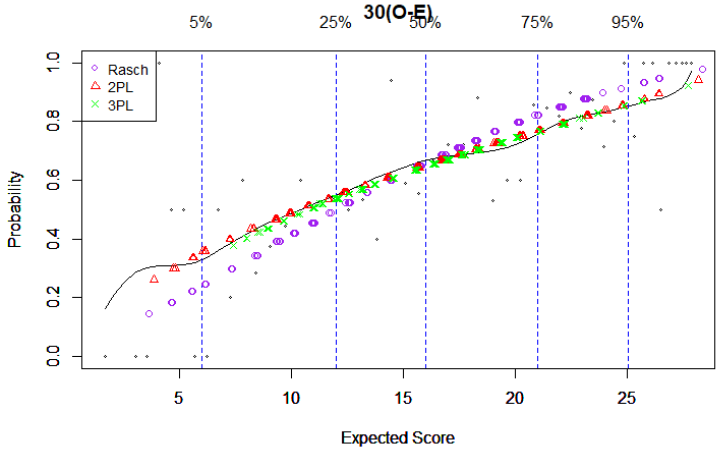


(c) Item 28: Out-of-sample-S

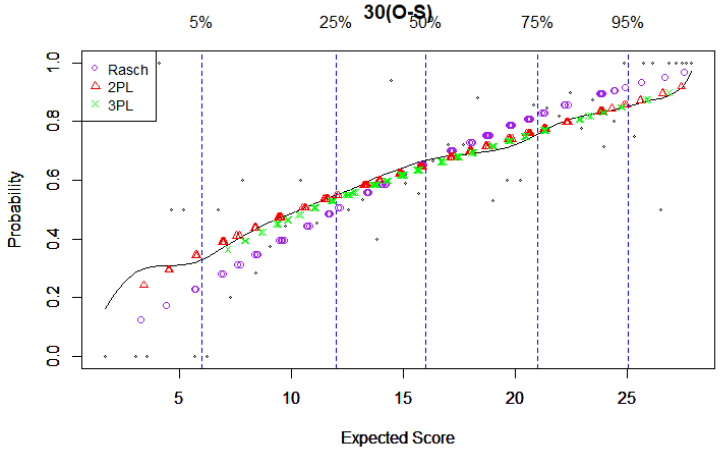
Figure D.28: Kernel Smoothing Checking Plot for Item 28.



(a) Item 30: In-sample



(b) Item 30: Out-of-sample-E



(c) Item 30: Out-of-sample-S

Figure D.29: Kernel Smoothing Checking Plot for Item 30.